Chapter

# Search for Human-Specific Proteins Based on Availability Scores of Short Constituent Sequences: Identification of a WRWSH Protein in Human Testis

*Shiho Endo, Kenta Motomura, Masakazu Tsuhako, Yuki Kakazu, Morikazu Nakamura and Joji M. Otaki*

## Abstract

Little is known about protein sequences unique in humans. Here, we performed alignment-free sequence comparisons based on the availability (frequency bias) of short constituent amino acid (aa) sequences (SCSs) in proteins to search for human-specific proteins. Focusing on 5-aa SCSs (pentats), exhaustive comparisons of availability scores among the human proteome and other nine mammalian proteomes in the nonredundant (nr) database identified a candidate protein containing WRWSH, here called FAM75, as human-specific. Examination of various human genome sequences revealed that FAM75 had genomic DNA sequences for either WRWSH or WRWSR due to a single nucleotide polymorphism (SNP). FAM75 and its related protein FAM205A were found to be produced through alternative splicing. The FAM75 transcript was found only in humans, but the FAM205A transcript was also present in other mammals. In humans, both FAM75 and FAM205A were expressed specifically in testis at the mRNA level, and they were immunohistochemically located in cells in seminiferous ducts and in acrosomes in spermatids at the protein level, suggesting their possible function in sperm development and fertilization. This study highlights a practical application of SCS-based methods for protein searches and suggests possible contributions of SNP variants and alternative splicing of FAM75 to human evolution.

**Keywords:** availability score, short constituent sequence (SCS), alternative splicing, single nucleotide polymorphism (SNP), testis, FAM75, FAM205A, human genome, human proteome

## 1. Introduction

The human species has unique traits among animals. It is well known that morphological and physiological traits such as erect bipedalism, speech and language, and long reproductive period are very different from those of other primate species. Only humans have high intelligence that fosters sophisticated

communications and complex societies. This intelligence is related to continuous brain development after birth in humans, which is not observed in other great apes, including chimpanzees [1]. The evolutionary emergence of these unique traits in humans likely contributes to human speciation. The simplest hypothesis to explain human uniqueness is that it originates from the uniqueness of constituent molecules (i.e., genes and proteins) themselves. In this "constituent hypothesis," humans have unique genes and proteins that do not exist in chimpanzees. A contrasting hypothesis is that constituent molecules are similar between humans and chimpanzees, but they are regulated differently in these species. That is, in this "regulatory hypothesis," a similar set of proteins may be produced but at different times (heterochrony), in different locations (heterotopy), in different amounts (heterometry), and in different usage (heterotypy) [2]. These regulatory changes in gene expression seem to be evolutionarily parsimonious and, indeed, are supported by comparative observations at phenotypic levels [3].

One line of support for the regulatory hypothesis comes from genomics and developmental expression studies. Following the announcement of a human genome release [4], the genomes of other great apes were sequenced [5–7]. Comparisons of DNA sequences between humans and chimpanzees have revealed that nucleotide differences are only 1.23% in aligned sequences, and most of these differences are thought to be functionally insignificant [5]. Further rigorous comparisons throughout these genomes have revealed that nucleotide differences are 4% and that they are mostly located in noncoding regions [8]. The expression patterns of some genes are different between humans and chimpanzees during development [9–12]. Differences in transcriptomes have revealed that species differences in expression patterns are tissue-dependent and that testes have the greatest difference [13, 14]. It has been speculated that the accumulation of small expression or regulatory differences leads to large phenotypic differences between humans and chimpanzees [14]. On the other hand, while these findings support the regulatory hypothesis, they do not necessarily reject the constituent hypothesis [15, 16]. RNA-mediated mechanisms for novel genes have been proposed together with the "out of the testis" hypothesis, in which testis is considered a tissue for experimenting with new genes [16]. Comparisons among transcriptomes in primates have revealed that many genes for spermatogenesis in testes, which likely inhibit apoptosis when mutated, are positively selected [17, 18].

Although these genome comparison studies advance this field, there are a few inherent problems. First, their results are heavily dependent on database quality because of their methodological nature. Most genome sequences were draft sequences at the time of public release, likely containing numerous sequencing and assembling mistakes. For example, the previous chimpanzee genome was assembled in reference to the human genome, which means that genomic regions in chimpanzees that are different from those in the human genome may have been assembled to create false sequences, although continuous revisions have been made [19]. Even in the human genome, many previous gene records generated by automated assemblers have been removed after revisions. Moreover, population sampling bias from the sequenced genome cannot be avoided when samples from a small number of individuals are sequenced. The case of a transcription factor, FOXP2 (forkhead box P2), is an object lesson: FOXP2 has been proposed to have played a key role in human-specific evolution by assisting speech and language [20], but that evidence is likely to be weak and probably incorrect because of sampling bias [21].

Second, such genome comparisons are largely based on sequence alignments [22, 23]. Although sequence alignment methods are powerful and probably the most important in comparison studies, sequences that do not contain relatively long regions of similarity cannot be compared well. In other words, short sequences that

do not extend to longer similarities are discarded as noise [22]. Although this strategy is highly successful, it assumes that nonaligned short sequences are not important, which may not always be true. There may still be important differences undiscovered where alignments are not possible.

An approach to the second issue above is to develop alignment-free methods. The advantage of the alignment-free approach is that any collections of proteins can be compared quantitatively. Although various types of alignment-free approaches have been developed [24, 25], including our previous attempts to use membrane topology [26] and a self-organizing map [27], the alignment-free approach in the present study is based on the "availability" (frequency bias) of short constituent sequences (SCSs) of amino acids (aa) in proteins [28–33]. The length of SCSs can be 2 aa (doublet), 3 aa (triplet), 4 aa (quartet), 5 aa (pentat), and more in a given protein. This SCS-based analysis is basically similar to other related analyses for amino acid sequence patterns that were called under different terms with slightly different mathematical operations: oligopeptide patterns [34–39], amino acid sequence repertoire [40], peptide vocabulary [41], *n*-gram [42, 43], *n*-tuple [44], and pseudo amino acid composition [45–47]. There are some noteworthy recent studies that encourage this line of approach: for example, nonrandom distributions of 5-aa SCS are demonstrated in the current proteome databases [38], confirming the previous finding that biological bias occurs in protein coding [28, 29]. Among these existing studies, our approach is operationally one of the simplest, and it emphasizes analogies between languages and protein sequences [32, 33]. Encouragingly, linguistic aspects of proteins have been noted in other studies [48, 49].

In our approach, protein sequences are considered to be composed of many SCSs. Importantly, the number of possible SCSs is limited because a protein is composed of just 20 kinds of amino acids; there are 400 ($=20^2$) permutations of 2-aa SCSs (doublets), 8000 ($=20^3$) permutations of 3-aa SCSs (triplets), 160,000 ($=20^4$) permutations of 4-aa SCSs (quartets), and 3,200,000 ($=20^5$) permutations of 5-aa SCSs (pentats). Frequencies of individual SCSs in a given protein database can be inferred theoretically based on frequencies of component amino acids, which is called the expected frequency ($E$). On the other hand, real frequencies of individual SCSs ($R$) in a given protein database can be obtained through database searches. The availability score ($A$) of a given SCS in a protein database can be simply defined as $A = (R - E)/E$. Availability scores thus indicate biological frequency bias that might have occurred for functional or historical reasons during protein evolution. In other words, availability scores ($A$) of SCSs are used instead of simple real frequencies ($R$) of SCSs to exclude noise from random occurrence.

Among *n*-SCSs, we state that 5-aa SCSs (pentats) are optimal for analyses for the following reasons [28, 29, 33]. First, they are practically manageable in number (exactly 3,200,000 different pentats) in our computational system. Higher computational power, which is sometimes not practical, is required to use 6-aa or longer SCSs. Second, the number of possible SCSs should be reasonably comparable to or smaller than the number of existing SCSs in a biological database. The use of 6-aa or longer SCSs would result in many nonexistent SCSs in the database because the number of possible 6-aa (or longer) SCSs is much larger than the number of existing SCSs in a given database. Third, 5-aa SCSs are likely structurally reasonable units (or "blocks") to build functional protein structures [50–54]. Fourth, it was suggested that small stretches of proteins are often recognized in protein interactions. For example, T-cell receptors recognize 5-aa SCSs as antigens in the process of antigen presentation, and this fact relates to the frequency bias of SCSs in parasites to avoid recognition by the T-cell receptors of the host [41]. Specificities of immune responses are thus likely influenced by SCSs in expressed proteins in a given organism, as also suggested by usage of rare SCSs as immune adjuvant vaccines [39].

Furthermore, rare SCS sequences evolved as untranslatable sequences in bacteria as a mean of translational control [40].

Using this simple concept of availability score, secondary structure characterization has been performed; SCS frequencies (and thus availability scores) are different among different secondary structures [30]. Availability scores are also different between parallel and antiparallel β-strands [31]. This approach is also relevant to identifying sequence motifs in some, although not all, proteins [32]. It has been shown that triplet compositions in proteomes may reflect phylogenetic relationships [32, 37, 53]. We believe that this approach is applicable to understanding species specificity.

We have implemented several applications as the SCS Package that informatically examine protein sequences [33]. Among them, we have built an application for identifying species-specific SCSs. In the present study, we compared human and other 9 mammalian proteomes based on availability analysis of 5-aa SCSs (pentats) to identify human-specific pentats. We hypothesized that a protein containing the identified human-specific pentat would be unique to humans and might have played a role in human evolution.

## 2. Materials and methods

### 2.1 The SCS package

Assuming that small changes in amino acids in proteins (or corresponding nucleotide changes in DNA) contribute significantly to phenotypic differences between humans and chimpanzees, the concept of SCS-based methods is to detect small amino acid usage differences between species in an alignment-independent manner. The SCS package is an open web service containing six applications (plus the latest application to analyze idiom networks under development [55]) for protein analyses (http://scspackage.ads.ie.u-ryukyu.ac.jp/) [33]. These applications run in reference to the database downloaded from the nonredundant (nr) database of the NCBI (National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda), which was downloaded on August 20, 2015. Because T-cell receptors, B-cell receptors, and antibodies are produced by somatic recombination and hypermutation, protein records containing the following keywords in sequence names were excluded: anti, IgG, IgM, IgA, IgD, IgE, BCR, TCR, B-cell receptor, T-cell receptor, Ig, and immunoglobulin. A complete match, including spaces, was required to be excluded. Frequencies and availability scores for all possible $n$-aa SCSs ($n$ = 3, 4, and 5 in the current SCS package) in the database were calculated and stored in the database [56]. For species comparison, each record in the downloaded database was sorted into its original species to produce species-specific proteome databases.

In this study, we focused on 5-aa SCSs (pentats). For multiple species comparison, the availability score difference, $\Delta A$, was calculated; for example, when a human 5-aa SCS had an availability score of 10 and the availability scores of that SCS in gorilla, pig, and mouse were 5, 3, and 2, respectively, the human $\Delta A$ was calculated as follows: $\Delta A = 10 \times 3 - (5 + 3 + 2) = 20$, where the multiplicative factor ($\times 3$) comes from the number of species to be compared. In this way, $\Delta A$ scores were assigned to all 3,200,000 pentats. The following nine species were used to obtain $\Delta A$ for human (*Homo sapiens*): chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo abelii*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), cow (*Bos taurus*), and pig (*Sus scrofa*).

### 2.2 Bioinformatics web services

After identifying FAM75 using the SCS package, available information on FAM75 and FAM205A was gathered using various web sites. The location FAM75/FAM205A on chromosomes and their single nucleotide polymorphism (SNP) variants were searched using Map Viewer (www.ncbi.nlm.gov/mapview/) in the NCBI server. For various information on human transcripts, we referred to H-InvDB (www.h-invitational.jp/hinv/ahg-db/index_ja.jsp) [57]. This site provides curated information on gene structure, splicing variants, functional RNAs, protein functions, functional domains, intracellular distribution, metabolic pathways, three-dimensional structures, disease relationships, genetic polymorphism (SNPs, indels, microsatellites, and others), gene expression profiles, molecular evolutionary characters, protein–protein interactions, and gene families. Tissue-specific expression profiles were searched using H-ANGEL (http://www.h-invitational.jp/hinv/h-angel/wge_top.cgi?), a database for human gene expression profiles, in the H-InvDB server. Information on alternative splicing variants of the nonhuman primates and mouse was obtained from Map Viewer in NCBI. We referred to the following latest annotations: chimpanzee (Annotation Release 103), western gorilla (Annotation Release 100), Sumatran orangutan (Annotation Release 102), and laboratory mouse (Annotation Release 106). We frequently used protein BLAST in the NCBI server for conventional similarity search (blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) [22] and performed multiple sequence alignments when necessary using MEGA7 (www.megasoftware.net/) [58]. In addition, the cDNA sequence of FAM75 was subjected to RegRNA analysis (regrna.mbc.nctu.edu.tw/html/about.html) to identify any possible sequence motifs in FAM75 mRNA [59].

To further examine SNP variants in human populations, we used dbSNP (www.ncbi.nlm.nih.gov/snp/) [60] and the 1000 Genome Project by IGSR (The International Genome Sample Resource) (www.internationalgenome.org) [61]. Protein expression was examined using The Human Protein Atlas (www.proteinatlas.org) [62, 63]. This site contains immunohistochemical data for various human tissues. For identification of transmembrane domains in FAM75 and FAM205A, SOSUI (harrier.nagahama-i-bio.ac.jp/sosui/) [64] and TMHMM (www.cbs.dtu.dk/services/TMHMM/) [65] were used. For the subcellular distributions of FAM75 and FAM205A, PSORT II Prediction (psort.hgc.jp/form2.html) [66] was used. Pfam (pfam.xfam.org) [67] was used for the identification of protein families. Two applications of the SCS Package, "sequence analysis based on availability scores of short constituent amino acid sequences" (scspackage.ads.ie.u-ryukyu.ac.jp/sequence-analysis.php) [32, 33] and "extraction of idiomatic connections between triplets in proteins" (scspackage.ads.ie.u-ryukyu.ac.jp/extraction-of--idiomatic-connections.php) [33], were used to identify possible functional sites in FAM75. EMBOSS Pepwindow (emboss.sourceforge.net/index.html) [68] was used for the Kyte-Doolittle hydropathy plot. These web sites were accessed mainly in 2017 and 2018 and were reconfirmed in 2019.

### 2.3 Human cDNA samples for tissue expression profiling

For the cDNA template, we purchased human MTC (multiple tissue cDNA) panels I and II (Takara Bio, Kusatsu, Shiga, Japan). The panels contain first-strand cDNA from polyA$^+$ RNA and are free from genomic DNA. The amounts of cDNA are approximately 1.0 ng/μL and are normalized to four housekeeping genes, phospholipase A2, G3PDH (glyceraldehyde-3-phosphate dehydrogenase), β-actin, and α-tubulin, which makes it possible to compare expression levels among different tissues. Panels I and II together contain cDNA samples from the following 16 human

tissues: heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, small intestine without mucosal lining, colon with mucosal lining, and peripheral blood leukocyte. Each tissue sample was pooled from 1 to 550 Caucasians, and the testis sample was pooled from 45 Caucasians aged 14–64, according to the manufacturer's specifications.

### 2.4 PCR primers

Based on the cDNA sequence of FAM75, we designed two sets of PCR primers for nested PCR using Primer-BLAST (www.ncbi.nlm.nih.gov/tools/primer-blast/). The first set was to amplify both FAM75 and FAM205A from the consensus region, and the second set was to amplify FAM205A from the region that is present only in FAM205A. For the first set, the first-round forward primer was 5'-TTACCAGG-TACTGTCACTGAACAC-3′, and its paired reverse primer was 5'-TTCTGAAGC-TAGACTCTGTAAGGC-3′. This first round of PCR was expected to amplify 1387 bp. The second-round (nested) forward primer was 5'-AGTTGTACA-GACGTTGCAAAAGAG-3′, and its paired reverse primer was 5'-TTTCTGAAGC-TAGACTCTGTAAGGC-3′. This second round (nested) PCR was expected to amplify 1097 bp.

For the second set, the first-round forward primer was 5'-ATATCCCTTATA-CATCTATGGCTCCATCTTC-3′, and its paired reverse primer was 5'-TTTTATTTCTGAAGCTAGACTCTGTAAGGC-3′. This round of PCR was expected to amplify 3608 bp. The second-round (nested) forward primer was 5'-GTATGTCTTTAGATCAGAGTCTGGAGTTTC-3′, and its paired reverse primer was 5'-TTTATTTCTGAAGCTAGACTCTGTAAGGCTG-3′. This round of PCR was expected to amplify 3206 bp.

### 2.5 PCR conditions

We used an Astec PC320 thermal cycler (Fukuoka, Japan) and Tks Gflex DNA polymerase (Takara Bio) for PCR. According to the manufacturer's specifications, this DNA polymerase has high fidelity; the error rate was reported to be 0.0131%.

The original cDNA sample from the human MTC panels (Takara Bio) was diluted 10 times to make PCR template samples. The following solutions were mixed to start PCR: Gflex PCR buffer 12.5 μL, deionized water 8.5 μL, DNA polymerase 0.5 μL, forward primer 0.5 μL, reverse primer 0.5 μL, and cDNA template 2.5 μL in a total amount of 25.0 μL. The nested PCR was performed using 2.5 μL reaction solution from the first-round PCR. In both the first and second (nested) rounds, a negative control was performed using deionized water without template cDNA.

The first PCR cycles were performed as follows: an initial denaturing step at 94°C (5 min); 10 cycles of 98°C (30 s), 60°C (30 s; −0.5°C/cycle), and 68°C (1 min); 30 cycles of 98°C (30 s), 55°C (30 s), and 68°C (1 min); and a last extension step at 68°C (30 s). The second (nested) PCR cycles were the same as the first PCR cycles except the duration of the initial denaturing step at 94°C (1 min). In both the first and second PCRs, the first 10 cycles were subjected to stepwise temperature reduction (i.e., touch-down PCR); the first cycle was 60.0°C, the second cycle was 59.5°C, and the third cycle was 59.0°C, and so on.

Positive controls were performed with G3PDH primers that were supplied in the Human MTC Panels (Takara Bio) from the manufacturer. The PCR product was expected to be 938 bp. The primer sequences were as follows: 5'-TGAAGGTCG-GAGTCAACGGATTTGGT-3′ for the forward primer and 5'-CATGTGGGCCAT-GAGGTCCACCAC-3′ for the paired reverse primer. PCR cycles were as follows: an

initial denaturing step at 95°C (1 min); 38 cycles of 95°C (30 s) and 68°C (3 min); and a final extension step at 68°C (3 min).

PCR products (1.0 μL) were subjected to 0.8% agarose gel electrophoresis in TAE buffer and stained with ethidium bromide for visualization. The PCR products were run with λHindIII DNA size marker (New England Biolabs, Ipswich, MA, USA).

## 3. Results

### 3.1 Identifying candidate human-specific pentats

Availability scores ($A$) were given to all possible pentats in the human proteome database. Among them, the top 10 pentats with the highest availability scores were HHHHH (rank 1; $A$ = 1837), MYGCD (rank 2; $A$ = 1770), MRYFY (rank 3; $A$ = 1321), WYWHF (rank 4; $A$ = 1262), PEYWD (rank 5; $A$ = 1140), MYQWW (rank 6; $A$ = 1100), HSMRY (rank 7; $A$ = 1096), NWHWA (rank 8; $A$ = 1041), WWNFG (rank 9; $A$ = 1007), and AWWNF (rank 10; $A$ = 928). Similarly, availability scores were given to all possible pentats in nine other mammalian proteome databases. Using the "extraction of species-specific amino acid sequences" program in the SCS Package, availability difference scores ($\Delta A$) were calculated for the human proteome. When pentats were ranked according to $\Delta A$ for humans, the top 10 pentats with the highest scores were MYGCD (rank 1; $\Delta A$ = 15,180), MRYFY (rank 2; $\Delta A$ = 11,777), WYWHF (rank 3; $\Delta A$ = 10,683), PEYWD (rank 4; $\Delta A$ = 9961), HSMRY (rank 5; $\Delta A$ = 9695), MYQWW (rank 6; $\Delta A$ = 9377), NWHWA (rank 7; $\Delta A$ = 9337), GQWRW (rank 8; $\Delta A$ = 8255), AWWNF (rank 9; $\Delta A$ = 7939), and EYWDR (rank 10; $\Delta A$ = 7878), showing similar but different rank orders from the human proteome alone. These pentats had large $\Delta A$ values, indicating that they are strongly preferred in human proteins.

Among the $\Delta A$ rank order of pentats, we focused on pentats that showed the lowest possible availability scores ($A = -1$) in all other nine mammalian proteome databases, meaning that these pentats did not exist in the nonhuman proteomes at all. We found WRWSH at rank 204 ($\Delta A$ = 1720) and MMFGC at rank 226 ($\Delta A$ = 1594) that met this criterion. However, MMFGC was found to be a false-positive, because this pentat was located exclusively in immunological proteins that could be subject to somatic recombination and hypermutation. Therefore, we decided to focus on WRWSH hereafter.

### 3.2 Human proteins containing WRWSH

Human proteins containing WRWSH were identified using the "search for amino acid sequences of species" program, one of the SCS Package programs. Among all 148 hits, 16 hits were related to "mucin-19-like isoform," 55 hits to "glycine-rich cell wall structural protein," 28 hits to "RNA-binding protein," 48 hits to "uncharacterized transmembrane protein," and 1 hit to "unnamed protein product." Unfortunately, these sequences except the last one, "unnamed protein product," were all "predicted" informatically as parts of "*Homo sapiens* Annotation Release 106" [69], and they were all removed from the latest annotation, "*Homo sapiens* Annotation Release 109" [70]. Because their status was uncertain at this point (although they resembled real protein sequences with a long open reading frame), they were not pursued in the present study. On the other hand, "unnamed protein product [*Homo sapiens*] (Accession No. BAC86357.1)", here called "FAM75"

based on the name of putative domain that it contained, was validated in the latest annotation [70], and thus, we pursued this protein for further investigation.

### 3.3 FAM75 and its related FAM205A

According to the NCBI record, FAM75 is a protein containing 1014 aa, and its cDNA coding sequence was 3274 bp (Accession No. AK125949.1). It is important to stress that FAM75 has been identified as cDNA from NEDO human cDNA sequencing project (www.nite.go.jp/en/nbrc/genome/project/annotation/cdna.html), and thus this protein is not likely an error product from genome sequencing. A protein BLAST search using FAM75 as a query identified the record "protein FAM205A [*Homo sapiens*] (Accession No. NP_001135389.1)." This protein record was closely related to the mRNA record "*Homo sapiens* family with sequence similarity 205 member A (FAM205A), mRNA (Accession No. NM_001141917.1)." The BLAST result showed that the identity score was 99%; 1003 aa were identical among 1014 aa. The record showed that FAM205A contained 1335 aa, and its cDNA coding sequence was 4311 bp. Thus, it was longer than FAM75. A DNA sequence comparison between FAM75 and FAM205A revealed that 16 bases were different (**Table 1**). When the FAM205A genomic DNA sequence (Accession No. NG_052658.1) was compared with its cDNA sequence, these 16 bases were identical (**Table 1**). Between FAM75 and FAM205A, 11 amino acids were different.

| FAM75 | | FAM205A | | FAM205A | |
|---|---|---|---|---|---|
| BAC86357.1, AK125949.1 | | NG_052658.1 | | NP_001135389.1, NM_001141917.1 | |
| (cDNA) | | (genomic DNA) | | (cDNA) | |
| 122 | T(S) | 8332 | C(P) | 1144 | C(P) |
| 139 | C(H) | 8349 | T(H) | 1161 | T(H) |
| 144 | C(S) | 8354 | T(E) | 1166 | T(E) |
| 584 | G(V) | 8794 | A(M) | 1606 | A(M) |
| 888 | C(S) | 9098 | T(L) | 1910 | T(L) |
| 894 | A(H) | 9104 | G(R) | 1916 | G(R) |
| 958 | C(G) | 9168 | T(G) | 1980 | T(G) |
| 1279 | A(P) | 9489 | G(P) | 2301 | G(P) |
| 1530 | T(V) | 9740 | A(E) | 2552 | A(E) |
| 1540 | T(P) | 9750 | C(P) | 2562 | C(P) |
| 2085 | T(I) | 10295 | G(S) | 3107 | G(S) |
| 2226 | A(E) | 10436 | G(G) | 3248 | G(G) |
| 2267 | T(Y) | 10477 | C(H) | 3289 | C(H) |
| 2391 | A(H) | 10601 | G(R) | 3413 | G(R) |
| 2582 | T(S) | 10792 | G(A) | 3604 | G(A) |
| 3145 | T | 11355 | C | 4167 | C |

Note: Numbers in this table indicate those of bases in the original records. Corresponding amino acids are shown in parenthesis. The shaded bases (and corresponding amino acids) correspond to the candidate human-specific pentat WRWSH in FAM75.

**Table 1.**
*Different DNA bases and protein amino acids between FAM75 (unnamed protein product) and FAM205A in the NCBI records.*

Interestingly, FAM205A in that record had WRWSR instead of WRWSH; the DNA sequences corresponding to the last amino acid of WRWS (H/R) were A (adenine) in FAM75 cDNA and G (guanine) in FAM205A cDNA and gDNA. These results suggest that the two products are closely related and may be produced from the same genomic site by alternative RNA splicing.

### 3.4 Gene structures: alternative splicing and polymorphism

A UniGene search revealed that the FAM7/FAM205A gene was located at 9p13.3 on chromosome 9 in the human genome [71]. As expected, their exon-intron structures were different (**Figure 1**). FAM75 had a single exon, whereas FAM205A had four exons. The exon of FAM75 had high homology with the fourth exon of FAM205A. The 5'-UTR of FAM75 also corresponded to the fourth exon of FAM205A. Clearly, these two RNA transcripts and their proteins are products of alternative splicing from the same genomic locus.

H-InvDB revealed two additional splicing variants (HIT000496944 and HIT000496575) from the same locus at 9p13.3 (**Figure 1**). The record HIT000496944 in the NCBI database was "*Homo sapiens* cDNA FLJ51393 complete code (AK302320.1)," and the record HIT000496575 was "*Homo sapiens* cDNA FLJ58301 complete code (AK301951.1)," both named "unnamed protein product." These are splicing variants, but among them, only FAM75 lacked the first 255-bp exon, indicating that the translation initiation sites are different in these mRNAs.

Because not all RNA transcripts are translated into proteins, we used a RegRNA search of UTRs (untranslated regions) to examine the integrity of the FAM75 mRNA. The RegRNA search revealed that the 5'-UTR of FAM75 had an internal ribosome entry site (IRES) [72–74] among other motifs, suggesting that the FAM75 mRNA is likely translated into proteins.

### 3.5 WRWSH and WRWSR in human populations

The G/A difference in FAM75/FAM205A in genomic DNA (corresponding to the H/R difference in WRWSH or WRWSR) was confirmed to be a SNP in humans, according to dbSNP. We found that this SNP was widespread in the human genome, and the G/A ratio was dependent on regional populations, as revealed by the 1000 Genomes Project (**Figure 2**). Among human populations, African populations had a high G frequency (i.e., WRWSR); the three highest G-frequency populations were Gambian in Western Division (96.16%); Yoruba in Ibadan,

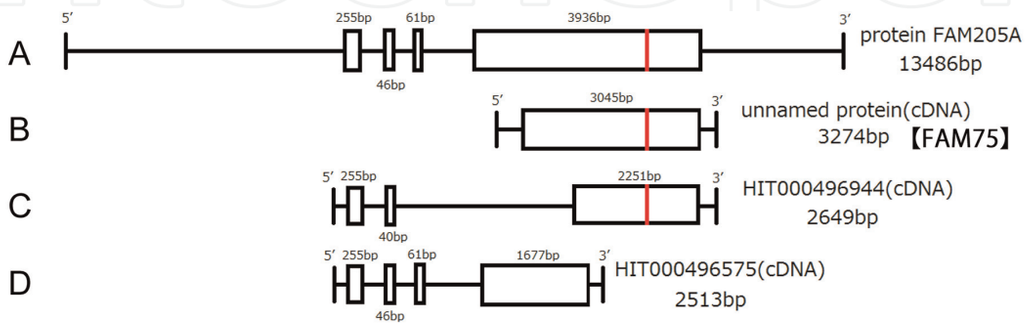


**Figure 1.**
*mRNA structures of FAM75, FAM205A, and their related transcripts from the same genomic locus in the human genome. (A) FAM205A from UniGene. (B) Unnamed protein (FAM75) from UniGene. (C) HIT000496944 from H-InvDB. (D) HIT000496575 from H-InvDB.*

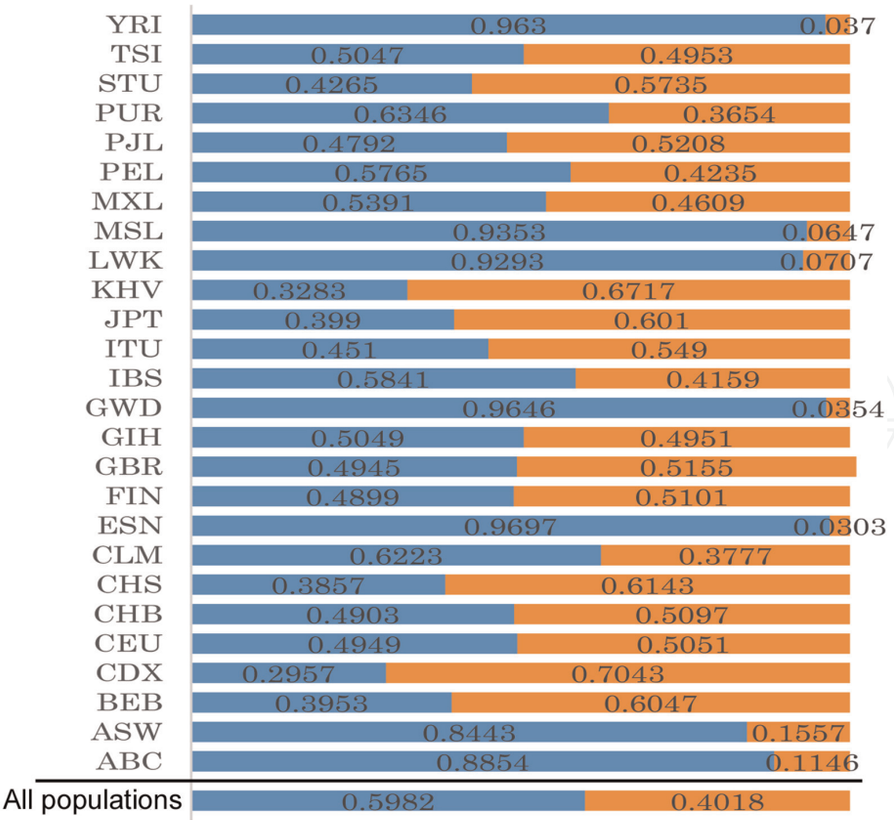


**Figure 2.**
*Genomic G/A ratio at the SNP site of the candidate human-specific pentat WRWSH in FAM75/FAM205A in various human populations. Abbreviations of populations or samples: YRI (Yoruba in Ibadan, Nigeria), TSI (Toscani, Italy), STU (Sri Lankan Tamil, UK), PUR (Puerto Rican, Puerto Rico), PJL (Punjabi in Lahore, Pakistan), PEL (Peruvian in Lima, Peru), MXL (Maxican ancestry in Los Angeles, CA, USA), MSL (Mende, Sierra Leone), LWK (Luhya in Webuye, Kenya), KHV (Kinh in Ho Chi Minh City, Vietnam), JPT (Japanese in Tokyo, Japan), ITU (Indian Telugu, UK), IBS (Iberian populations, Spain), GWD (Gujarati Indians in Houston, TX, USA), GBR (British from England and Scotland), FIN (Finnish, Finland), ESN (Esan, Nigeria), CLM (Colombian in Medellín, Colombia), CHS (Han Chinese south, China), CHB (Han Chinese in Beijing, China), CEU (Utah residents (CEPH) with northern and Western European ancestry), CDX (Chinese Dai in Xishuangbanna), BEB (Bengali, Bangladesh), ASW (African ancestry in SW, USA), and ABC (African Caribbean, Barbados).*

Nigeria (96.30%); and Mende in Sierra Leone (93.53%). In contrast, Asian and European populations had relatively high A frequency (i.e., WRWSH); the three highest A-frequency populations were Chinese Dai in Xishuangbanna (70.43%); Kinh in Ho Chi Minh City, Vietnam (67.17%); and Han Chinese South, China (61.43%).

### 3.6 Homologous proteins and alternative splicing products in other animals

Here, we searched for homologous proteins for FAM75 and FAM205A in other animals. Among the nine mammals used for the initial identification for WRWSH, homologous proteins for FAM205A were identified by BLAST search (**Table 2**); all proteins were FAM205A homologs, and all proteins in primates (chimpanzee, gorilla, and orangutan) contained WRWSR (corresponding to FAM205A) but not WRWSH (corresponding to FAM75), suggesting that WRWSH in FAM75 may be unique in humans among primates. In other nonprimate animals that were examined here, this pentat sequence was either not conserved at all or nonexistent.

To further examine whether splicing variants exist in other great apes, we checked the genome loci and transcript data using Map Viewer. In chimpanzees (**Figure 3**) and gorillas (not shown), there were no alternative splicing transcripts from this locus. In orangutans (not shown), there were three isoforms, the X1, X2,

| | Name in FASTA file | ID | Pentat |
|---|---|---|---|
| Human | unnamed protein | BAC86357.1 | WRWSH |
| Human | protein FAM205A | NG_052658.1 | WRWSR |
| Chimpanzee | protein FAM205A | XP001164235.2 | WRWSR |
| Gorilla | protein FAM205A like | XP_004048025.1 | WRWSR |
| Orangutan | protein FAM205A isoform | XP_009242592.1 | WRWSR |
| Mouse | predicted gene 12,429 isoform | XP_011248363.1 | SLQAQ |
| Rat | protein FAM205-A isoform | XP_008774156.1 | SQQGH |
| Opossum | protein FAM205-A like isoform | XP_007498908.1 | HVGNR |
| Platypus | protein FAM205A | XP_007657228.1 | ::::: |
| Cow | protein FAM205A | XP_001253501.1 | WQRRH |
| Pig | — | — | — |

*Note: Amino acid sequences are conceptual translation from genomic data. Red letters indicate amino acids different from those of the human pentat WRWSH. No corresponding pentat was found in the platypus (:::::), and no homologous protein was found in the pig (—).*

**Table 2.**
*Amino acid pentat sequences from mammals that are homologous to the human WRWSH in FAM75.*

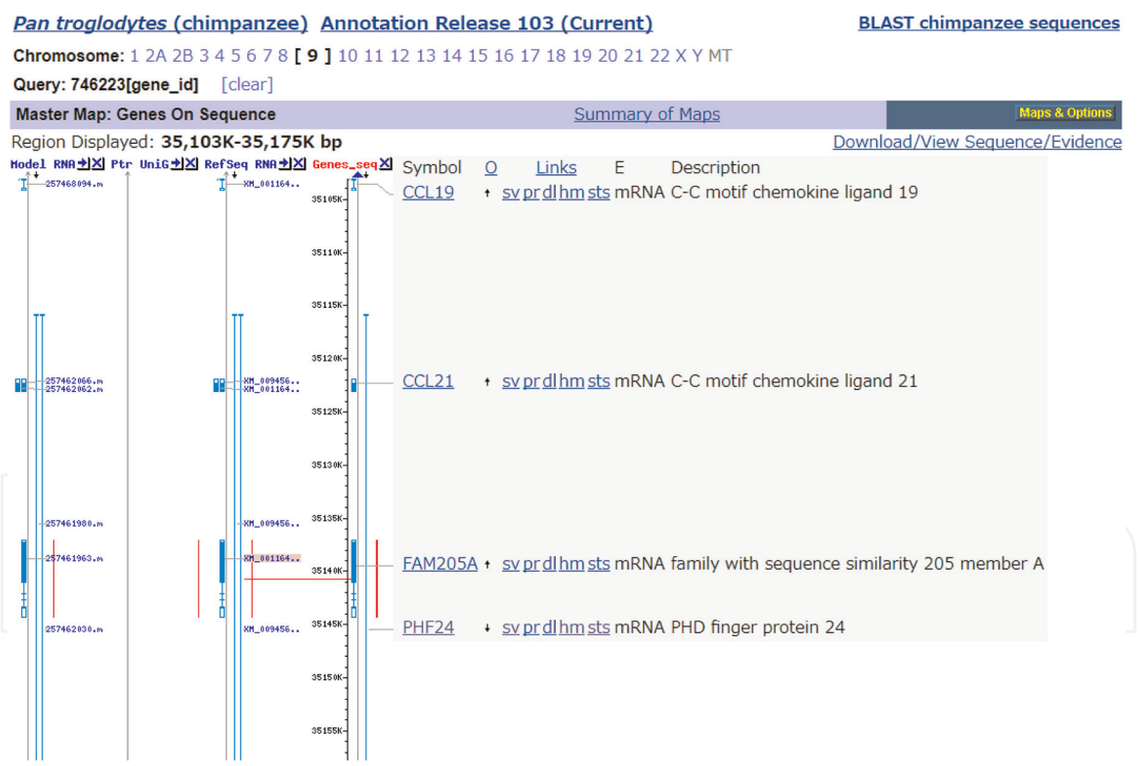


**Figure 3.**
*FAM205A and its surrounding locus of chromosome 9 in the chimpanzee genome.*

and X3 transcripts, from this locus. However, these transcripts were very similar to one another, and they were all considered FAM205A homologs containing WRWSR. We also examined the genome of the mouse as a representative nonprimate mammal (not shown). There were three transcript variants: "predicted gene 12429 isoform X1, X2" and "predicted gene 12429." They all contained SLQAQ instead of WRWSH in these proteins, and their splicing patterns were different from those of FAM75. We confirmed that human splicing patterns (**Figure 4**) were

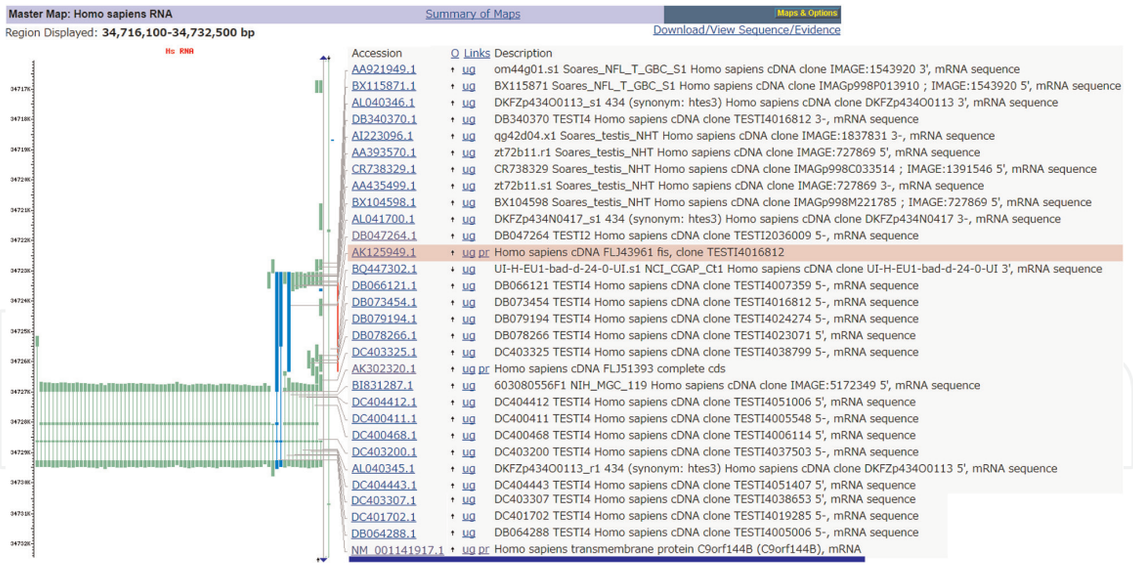


**Figure 4.**
*FAM205A/FAM75 and its surrounding locus of chromosome 9 in the human genome. FAM75 is highlighted in pink, and FAM205A is underlined in blue.*

different from those of these mammals. Therefore, we conclude that the FAM75 transcript was found only in humans.

### 3.7 Testis-specific expression of FAM75 and FAM205A

To examine its existence and expression in our laboratory, we performed RT-PCR (reverse transcription polymerase chain reaction) using two sets of PCR primers using 16 different human-tissue cDNA pools as templates. The first set of primers was designed to amplify both FAM75 and FAM205A (**Figure 5A**), and the second set was designed to amplify FAM205A only (**Figure 5B**). Due to their overlapping nature, exclusive amplification of FAM75 was not possible. In both primer sets, testis-specific expression was observed. A positive control using a primer set for G3PDH showed amplification from all tissues (**Figure 5C**), and a negative control (without cDNA template but with experimental primer sets) did not show any amplification.

Our results were consistent with the H-ANGEL expression database; in this database, FAM75 and FAM205A were not differentiated, but the database indicated that the expression was testis-specific (**Figure 6**). The NCBI database also indicated the testis-specific expression of FAM205A (not shown). The expression pattern of FAM205A was also found in the Human Protein Atlas, in which FAM205A was expressed in testis and in no other tissues examined at the mRNA level (not shown), confirming our PCR-based data. According to the Human Protein Atlas, cells in the seminiferous ducts (sperm and immature sperm cells) of the testis were clearly detected, but Leydig cells were not stained immunohistochemically (**Figure 7**). As mentioned in the Human Protein Atlas, staining was clearly detected in acrosomes in spermatids (**Figure 7**). Considering that the antibody used in the Human Protein Atlas could not differentiate FAM205A and FAM75 (because a recombinant C-terminal 104 aa fragment that is almost identical in both FAM205A and FAM75 was used as an antigen), both proteins were likely stained in the tissue sections.

### 3.8 Structural and functional predictions

We performed several sequence analyses to characterize the sequences of FAM75 (**Figure 8**). When FAM75 and FAM205A were subjected to SOSUI, the

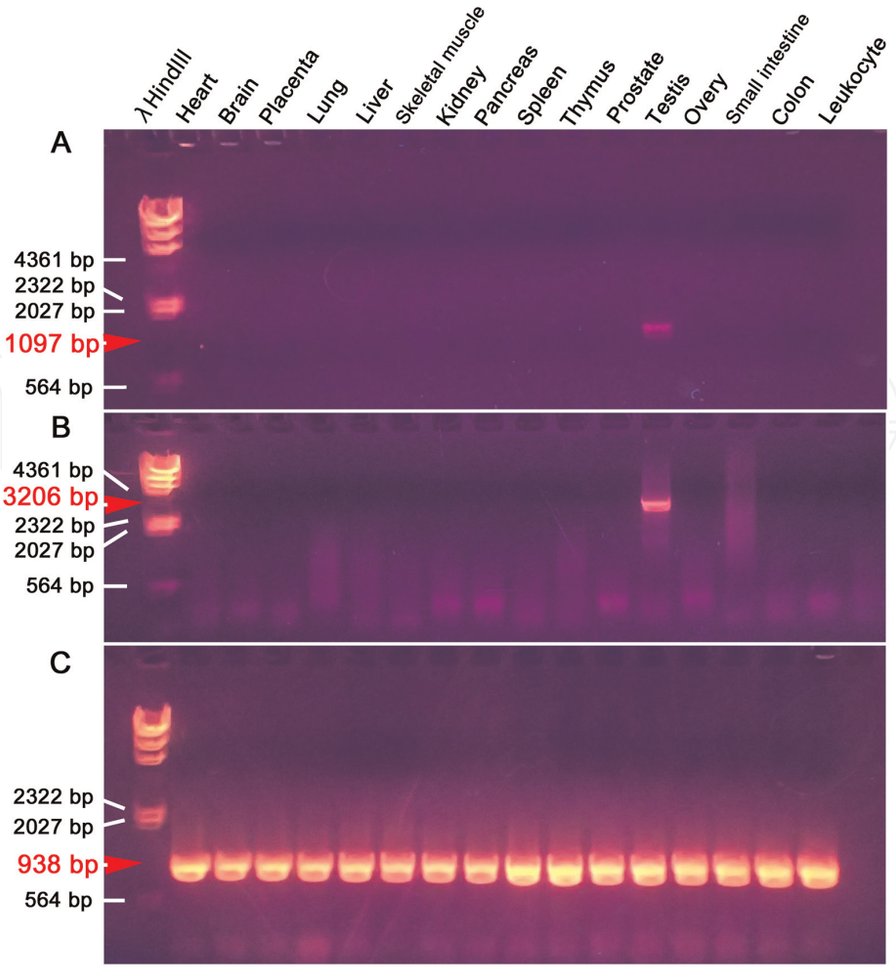


**Figure 5.**
*PCR products from human-tissue cDNA templates. (A) FAM75 and FAM205A. Primers were designed to amplify both FAM75 and FAM205A. A DNA fragment with the expected size (1097 bp) was amplified only from testis. (B) FAM205A. Primers were designed to amplify only FAM205A. A DNA fragment with the expected size (3206 bp) was amplified only from testis. (C) G6PDH as a positive control. A DNA fragment with the expected size (938 bp) was amplified from all tissues tested.*

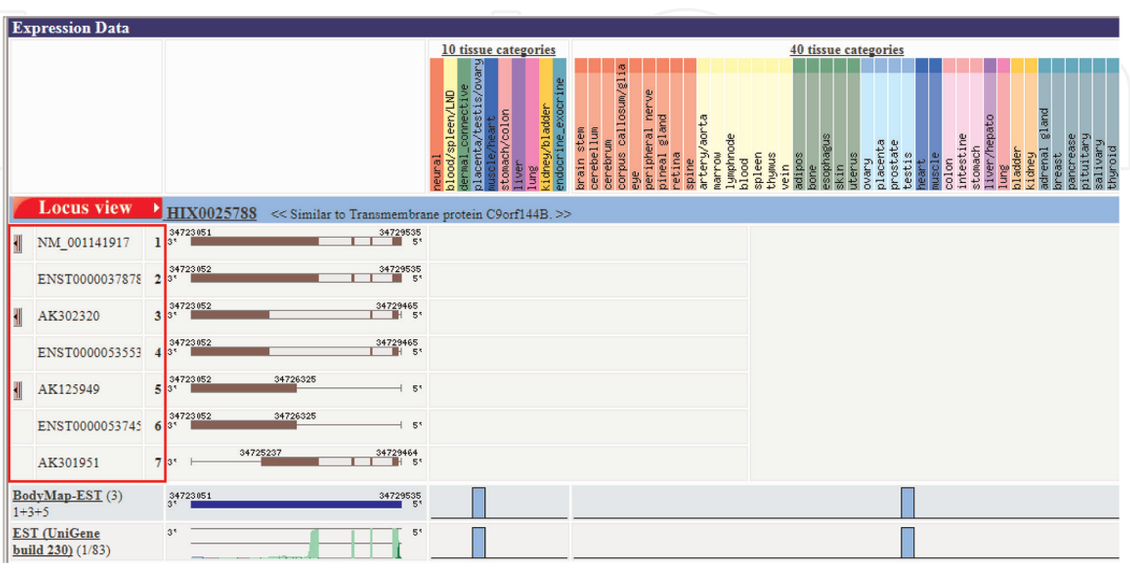


**Figure 6.**
*Gene expression profile of FAM205A in human tissues (H-ANGEL). NM_001141917 indicates FAM205A, and AK125949 indicates FAM75. http://www.h-invitational.jp/hinv/h-angel/wge_server.cgi?gpid=HIX0025788.*

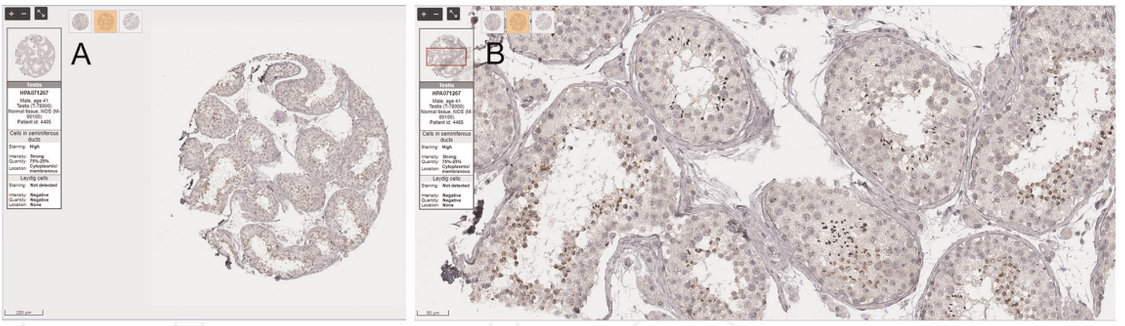


**Figure 7.**
*Immunohistochemical detection of FAM205A in a normal human testis section. FAM75 is also likely stained if it is present, because the antibody was raised against a recombinant C-terminal 104-aa fragment, which is found both in FAM205A and FAM75. Pictures were taken from the Human Protein Atlas (HPA071267, male, age 41, testis (T-78000), normal tissue, NOS (M-00100), patient ID: 4485). Dark brown signals are seen in spermatids and other differentiating cells in the seminiferous ducts. The crescent-like staining likely represents developing acrosomes. Two additional samples (ages 25 and 65) are shown in the Human Protein Atlas with essentially the same results. (A) Entire section. (B) High magnification of A.*

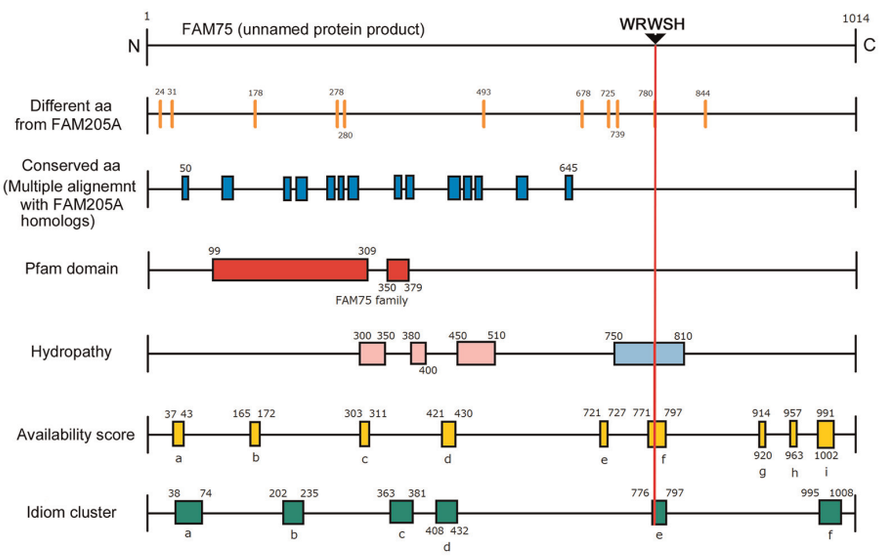


**Figure 8.**
*Identification of possible functional sites in FAM75. Shown are different amino acids from FAM205A (NCBI GenBank record), conserved amino acids based on multiple alignment with FAM205A homologs, Pfam domain, hydropathy, availability score, and idiom clusters. The location of the amino acid H in the candidate human-specific pentat WRWSH is indicated by a vertical red line.*

former was predicted as a soluble protein, but the latter was predicted as a membrane protein with a single transmembrane helix. TMHMM also showed essentially the same results. Indeed, the Human Protein Atlas considered FAM205A to be both a membrane protein and a cytoplasmic protein based on immunohistochemical results, suggesting that FAM75 and FAM205A may be detected in the cytoplasm and in membranes, respectively, as predicted by SOSUI and TMHMM. In contrast, both were predicted to be "nuclear" using PSORT II Prediction.

To search for possible functional sites, different amino acids between FAM75 and FAM205A (**Table 1**), conserved amino acids among FAM205A and similar sequences (top 100 BLAST data) based on multiple alignment, Pfam domain data, a hydropathy plot, an availability plot, and an idiom plot were aligned together (**Figure 8**). Conserved amino acids were located mostly in the N-terminal side, in which the "FAM75 domain" identified by Pfam was also located. WRWSH was located at the center of the hydrophilic region in the C-terminal side and

corresponded to a high availability region and a high idiom region, although their significance was not clear at this point.

## 4. Discussion

In this paper, we identified a WRWSH-containing protein, FAM75, as a candidate human-specific protein. We assumed that pentats with high availability scores in humans and no occurrence ($A = -1$) in nine other mammals might be contained in a human-specific protein. The current method based on this assumption indeed identified FAM75. Although the DNA sequence coding for WRWSH is one of the SNP variants in the human genome (i.e., WRWSH was not conserved in all human populations), this fact does not exclude the candidacy of WRWSH as a human-specific pentat, because we do not know when this SNP variant was created during human history. Likely, not all point mutations are functionally equal; some point mutations may incidentally create a rare pentat like WRWSH that may contribute to functional novelty. Interestingly, the FAM75 transcript was found only in humans as an alternative splicing transcript of FAM205A. In this sense, our SCS-based search for human-specific proteins successfully identified what we wanted to identify. The success of this study may simply be fortunate. On the other hand, there are many other candidate human-specific pentats that we did not examine in detail. Changing search conditions, including the length of amino acid sequences (i.e., triplets, quartets, and longer SCSs), could identify further candidate human-specific SCSs.

The present study showed that the SCS-based approach is a relevant addition to a list of practical sequence comparison methods. As with other methods, the SCS-based method is influenced by SNPs, accuracy, and the amount of information in databases. For example, the human genome has numerous SNP variations, and there is much less genomic information for other primates than for humans. $A$ and $\Delta A$ scores, which were used to search in this study, are dependent on databases. WRWSH had high $\Delta A$ between humans and nine other mammals, and this is partly because there were many human protein records that contained this pentat at the time of the database search. Unfortunately, most of these records were later removed from human databases (NCBI GenBank records) because of the uncertainty of their status (although they were not rejected completely). This illustrates the importance of database quality in genome comparison studies. However, whatever $\Delta A$ was, we focused on the pentats that were not used at all ($A = -1$) in the nine other nonhuman mammals, which made the choice of pentats for further investigation less sensitive to database quality.

FAM75 and FAM205A appear to be alternative splicing products from the same genomic locus in humans (**Figure 1**). The relationship of the evolutionary invention of FAM75 as an alternative splicing product with that of a SNP variant for WRWSH is unclear. We cannot exclude the possibility that this may be a simple coincidence, but this coincidence is in accordance with our starting hypothesis for this study: proteins containing a human-specific pentat may indeed be human-specific as proteins. We confirmed the expression of FAM205A and/or FAM75 at the mRNA level in human tissues (**Figure 5**). At the protein level, the FAM205A protein (and probably also the FAM75 protein) was shown to be located in cells in seminiferous ducts and in acrosomes in spermatids in the testis (**Figure 7**). Interestingly, FAM205A was also detected in the human sperm nucleus in a proteomic study [75]. Although it is difficult to distinguish FAM75 and FAM205A at the mRNA and protein levels, it is demonstrated that the FAM75/FAM205A gene is not a pseudogene, and protein products are actively produced in testis. The discovery of

the IRES element in FAM75 mRNA also supports the idea that FAM75 mRNA is actively translated into proteins. On the other hand, we found two additional alternative splicing products in H-InvDB (**Figure 1**). These additional mRNAs were not examined in this paper, because of insufficient information. However, their status is of interest if they really exist; they may have similar but slightly different functions from FAM205A and FAM75.

Mechanistically, alternative splicing may be a relatively easy way to create a new protein sequence. It may be considered not only a "regulatory change" (according to the regulatory hypothesis, because the evolutionary invention of a new alternative splicing product conserves the original protein-coding DNA sequence and gene function and thus is more conservative with respect to species evolution) but also a "sequence change" (according to the constituent hypothesis, because the protein sequence is changed). These two modes are likely intermingled in this case. To extrapolate this argument, transcriptome studies of alternative splicing or RNA processing may be fruitful to identify human-specific genes. The present discovery of the IRES element in the FAM75 mRNA may be surprising because IRES elements are mostly viral, and cellular elements are relatively rare [72–74]. A search for IRES elements in the genome may also be fruitful.

The evolution of WRWSH and FAM75 in relation to human speciation is an important but uncertain aspect to be discussed. There are two kinds of "human-specific" proteins. First, a group of proteins may have been involved in the early step of speciation of *Homo sapiens* from its ancestral species. Second, after the establishment of *Homo sapiens*, additional changes in a group of proteins may occur as a reinforcement process. In either case, these proteins may be called human-specific. If the pentat WRWSH (or FAM75) played a role in these early or late steps of human speciation, this pentat is human-specific, and it would be later mutated back to WRWSR in African populations. In this case, WRWSH was once assimilated completely in the human population during speciation, and a new WRWSR sequence is now assimilating, as genetic assimilation has been considered a key process in species evolution [76–78]. However, because WRWSH is relatively rare in African populations, it is more parsimonious to think that WRWSH evolved after human speciation in Asian or European populations. We speculate that FAM75 may have been invented from FAM205A to play a role in human speciation, but at least in the early stage, FAM75 exclusively contained WRWSR, as in the other great apes. WRWSH may then have been invented in FAM75 to reinforce human speciation. Alternatively, WRWSH did not play any role in human speciation, and its reinforcement simply fortified the function of FAM75 in some populations relatively recently.

What is the function of FAM75 in human testes? According to the results of immunohistochemistry (the Human Protein Atlas), SOSUI, and TMHMM, we speculate that FAM75 appears to function differently from FAM205A in different cellular sites. Because FAM75 is likely located in acrosomes (**Figure 7**), this protein may be involved in the process of fertilization. A possibility is that FAM75 confers human specificity to prevent cross-species fertilization with ancestral species. The FAM75/FAM205A genomic locus in humans has an additional two alternative splicing products, which were not pursued in the present study, and orangutans and mice appeared to have three transcripts from the same locus. It is tempting to speculate that this locus partly contributes to speciation in primates and other mammals by restricting cross-species fertilization in ancestral species.

Molecularly, the main function of FAM75 may be located in the "FAM75 domain" located at the N-terminal side of the molecule (**Figure 8**), but because WRWSH is located in a hydrophilic region at the C-terminal side of the molecule, this hydrophilic region may function in human specificity. Indeed, the conserved

regions are mostly located at the N-terminal side, probably for the general function of FAM75. The hydrophilic region also coincides with high availability and idiom-cluster regions.

Testis is known to be the tissue of the fastest evolution among other tissues based on gene expression comparisons in mammals, including the great apes [13, 14, 16–18, 79]. This flexibility may reflect diverse species-specific sexual behaviors. Mating is nonselective and frequent in chimpanzees, and only the highest-ranked male can mate in gorillas [80]. These behaviors have been thought to be related to testis-size differences; the chimpanzee has relatively large testes, and the gorilla has small ones [80]. Human testis size lies between these extremes, which may be related to the molecular evolution of FAM75 to modulate sperm development in testes or to withstand moderate sperm competition.

A recent finding that the gene locus for FAM205A is a susceptible locus for intracerebral hemorrhage (ICH) [81] is somewhat surprising. Either FAM205A or FAM75 may be expressed in cerebral cells at low levels or in restricted regions of the brain. It is tempting to speculate that a pleiotropic protein for both fertilization and brain development, such as FAM75/FAM205A, might have played a role in human evolution. The fact that the FAM205A/FAM75 gene is located not in a sex chromosome but in chromosome 9, despite its expression in the testis, might further suggest its dual role in sexual and nonsexual aspects of human specificity.

## 5. Conclusions

Our SCS-based approach identified FAM75, a WRWSH-containing protein, as a candidate human-specific protein. Its uniqueness in humans may be acquired not only by a point mutation for WRWSH but also by novel alternative splicing. Together with FAM205A, FAM75 is likely expressed in human testis, and its possible expression in acrosomes suggests its potential function in fertilization and thus in human speciation. Its potential pleiotropic function in the brain is very interesting and may also be investigated in the future.

## Acknowledgements

We thank Miki Kawauchi, Motosuke Tsutsumi, Hideka Konno, and other members of the BCPH Unit of Molecular Physiology for technical assistance and discussions. This work was supported by the Sekisui Chemical Grant Program for Research to JMO. This work was also supported by basic funds to JMO and MN from the University of the Ryukyus.

## Conflict of interest

Authors declare no competing interests.

## Author details

Shiho Endo[1], Kenta Motomura[2], Masakazu Tsuhako[1], Yuki Kakazu[2], Morikazu Nakamura[2] and Joji M. Otaki[1*]

1 The BCPH Unit of Molecular Physiology, Department of Chemistry, Biology and Marine Science, Faculty of Science, University of the Ryukyus, Okinawa, Japan

2 Department of Information Science, Faculty of Engineering, University of the Ryukyus, Okinawa, Japan

*Address all correspondence to: otaki@sci.u-ryukyu.ac.jp

## References

[1] Leigh SR. Brain growth, life history, and cognition in primate and human evolution. American Journal of Primatology. 2004;**62**:139-162

[2] Gilbert SF, Epel D. Ecological Developmental Biology. 2nd ed. Sunderland, MA: Sinauer Associates; 2015

[3] King M-C, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975;**188**:107-116

[4] Lander ES et al. Initial sequencing and analysis of the human genome. Nature. 2001;**409**:860-921

[5] Mikkelsen T et al. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;**437**:69-87

[6] Locke DP et al. Comparative and demographic analysis of orangutan genomes. Nature. 2011;**469**:529-533

[7] Scally A et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;**483**:169-175

[8] Vark A, Geschwing DH, Eichler EE. Explaining human uniqueness: Genome interactions with environment, behaviour, and culture. Nature Review Genetics. 2008;**9**:749-763

[9] McLean CY et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature. 2011;**471**:216-219

[10] Enard W et al. Intra- and interspecific variation in primate gene expression patterns. Science. 2002;**296**: 340-343

[11] Preuss TM, Caceres M, Oldham MC, Geschwind DH. Human brain evolution: Insights from microarrays. Nature Review Genetics. 2004;**5**:850-860

[12] Boyd JL, et al. Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. Current Biology. 2015;**25**: 772-779

[13] Khaitovich P, et al. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science. 2005;**309**: 1850-1854

[14] Khaitovich P, Enard W, Lachmann M, Pääbo S. Evolution of primate gene expression. Nature Review Genetics. 2006;**7**:693-702

[15] Nshon J-L. Birth of 'human-specific' genes during primate evolution. Genetica. 2003;**118**:193-208

[16] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Research. 2010;**20**:1313-1326

[17] Nielsen R, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biology. 2005;**3**:e170

[18] Tay SK, Blythe J, Lipovich L. Global discovery of primate-specific genes in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2009;**106**: 12019-12024

[19] Kronenberg ZN et al. High-resolution comparative analysis of great ape genomics. Science. 2018;**360**:eaar6343

[20] Enard W, et al. Molecular evolution of FOXP2, a gene involved in speech and language. Nature. 2002;**418**:869-872

[21] Atkinson EG, et al. No evidence for recent selection at *FOXP2* among diverse human populations. Cell. 2018; **174**:1424-1435.e15

[22] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;**215**:403-410

[23] Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: Algorithms for computing spliced alignments with identification of paralogs. Biology Direct. 2008;**3**:20

[24] Vinga S, Almeida JS. Alignment-free sequence comparison—A review. Bioinformatics. 2003;**19**:513-523

[25] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biology. 2017;**18**:186

[26] Otaki JM, Firestein S. Length analyses of mammalian G-protein-coupled receptors. Journal of Theoretical Biology. 2001;**211**:77-100

[27] Otaki JM, Mori A, Itoh Y, Nakayama T, Yamamoto H. Alignment-free classification of G-protein-coupled receptors using self-organizing maps. Journal of Chemical Information and Modeling. 2006;**46**:1479-1490

[28] Otaki JM, Ienaka S, Gotoh T, Yamamoto H. Availability of short amino acid sequences in proteins. Protein Science. 2005;**14**:617-625

[29] Otaki JM, Gotoh T, Yamamoto H. Potential implications of availability of short amino acid sequences in proteins: An old and new approach to protein decoding and design. Biotechnology Annual Review. 2008;**14**: 109-141

[30] Otaki JM, Tsutsumi M, Gotoh T, Yamamoto H. Secondary structure characterization based on amino acid composition and availability in proteins. Journal of Chemical Information and Modeling. 2010;**50**:690-700

[31] Tsutsumi M, Otaki JM. Parallel and antiparallel β-strands differ in amino acid composition and availability of short constituent sequences. Journal of Chemical Information and Modeling. 2011;**51**:1457-1464

[32] Motomura K, Fujita T, Tsutsumi M, Kikuzato S, Nakamura M, Otaki JM. Word decoding of protein amino acid sequences with availability analysis: A linguistic approach. PLoS One. 2012;**7**: e50039

[33] Motomura K, Nakamura M, Otaki JM. A frequency-based linguistic approach to protein decoding and design: Simple concepts, diverse applications, and the SCS package. Computational and Structural Biotechnology Journal. 2013;**5**: e201302010

[34] Bresell A, Persson B. Characterization of oligopeptide patterns in large protein sets. BMC Genomics. 2007;**8**:346

[35] Tuller T, Chor B, Nelson N. Forbidden penta-peptides. Protein Science. 2007;**16**:2251-2259

[36] Figureau A, Soto MA, Tohá J. A pentapeptide-based method for protein secondary structure prediction. Protein Engineering. 2003;**16**:103-107

[37] Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS. Proteomic signatures: Amino acid and oligopeptide compositions differentiate among phyla. Proteins. 2004;**54**:20-40

[38] Poznański J, et al. Global pentapeptide statistics are far away from expected distributions. Scientific Reports. 2018;**8**:15178

[39] Patel A, et al. Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. PLoS One. 2012;**7**: e43802

[40] Navon SP, et al. Amino acid sequence repertoire of the bacterial proteome and the occurrence of untranslatable sequences. Proceedings of the National Academy of Sciences of the United States of America. 2016;**113**: 7166-7170

[41] Zemková M, Zahradnik D, Mokrejš M, Flegr J. Parasitism as the main factor shaping peptide vocabularies in current organisms. Parasitology. 2017;**144**:975-983

[42] Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports. 2017;**7**:12961

[43] Vries JK, Liu X, Bahar I. The relationship between n-gram patterns and protein secondary structure. Proteins. 2007;**68**:830-838

[44] Daeyaert F, Moereels H, Lewi PJ. Classification and identification of proteins by means of common and specific amino acid n-tuples in unaligned sequences. Computer Methods and Programs in Biomedicine. 1998;**56**:221-233

[45] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins. 2001;**43**: 246-255

[46] Chou KC, Cai YD. Predicting protein quaternary structure by pseudo amino acid composition. Proteins. 2003; **53**:282-289

[47] Cai YD, Chou KC. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. Journal of Proteome Research. 2005;**4**:967-971

[48] Popov O, Segal DM, Trifonov EN. Linguistic complexity of protein sequences as compared to texts of human language. BioSystems. 1996;**38**:65-74

[49] Eroglue S. Language-like behavior of protein length distribution in proteomes. Complexity. 2014;**20**:12-21

[50] de Brevern AG, Valadié H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. Protein Science. 2002;**11**:2871-2886

[51] de Brevern AG. New assessment of a structural alphabet. In Silico Biology. 2005;**5**:283-289

[52] Joseph AP, et al. A short survey on protein blocks. Biophysical Reviews. 2010;**2**:137-145

[53] de Brevern AG, Joseph AP. Species specific amino acid sequence-protein local structure relationships: An analysis in the light of a structural alphabet. Journal of Theoretical Biology. 2011;**276**: 209-217

[54] Nekrasov AN, et al. A minimum set of stable blocks for rational design of polypeptide chains. Biochimie. 2019; **160**:88-92

[55] Kakazu Y, Nakamura M, Otaki JM. Idiom networks for short constituent sequences of amino acids. In: 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). 2005. pp. 15-19

[56] Kakazu Y, Nakamura M, Otaki JM. GPU acceleration for availability scoring of short constituent amino acid sequences. In: 2015 Third International Symposium on Computing and Networking (CANDAR). 2015. pp. 598-600

[57] Takeda J, et al. H-InvDB in 2013: An omics study platform for human functional gene and transcript discovery. Nucleic Acids Research. 2013;**41**:D915-D919

[58] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;**33**:1870-1874

[59] Huang HY, Chien CH, Jen KH, Huang HD. RegRNA: A regulatory RNA motifs and elements finder. Nucleic Acids Research. 2006;**34**: W429-W423

[60] Kitts A, Sherry S. The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. In: McEntyre J, Ostell J, editors. The NCBI Handbook. Bethesda: National Center for Biotechnology Information. Chapter 5; 2002

[61] Sudmant PH et al. An integrated map of structural variation in 2504 human genomes. Nature. 2015;**526**:75-81

[62] Fagerberg L et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular & Cellular Proteomics. 2014;**13**:397-406

[63] Uhlén M et al. Tissue-based map of the human proteome. Science. 2015;**347**: 1260419

[64] Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: Classification and secondary structure prediction system for membrane proteins. Bioinformatics. 1998;**14**:378-379

[65] Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model; application to complete genomes. Journal of Molecular Biology. 2001;**305**:567-580

[66] Nakai K, Horton P. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends in Biochemical Sciences. 1999;**24**:34-35

[67] El-Gebali S, et al. The Pfam protein families database in 2019. Nucleic Acids Research. 2018;**47**:D427-D432

[68] Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. Trends in Genetics. 2000;**16**:276-277

[69] NCBI. *Homo sapiens* Annotation Release 109. Date of submission of annotation to the public databases: March 26, 2018. Available at: www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/109/

[70] NCBI. *Homo sapiens* Annotation Release 106. Date of submission of annotation to the public databases: February 3, 2014. Available at: www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/106/

[71] Humphray SJ et al. DNA sequence and analysis of human chromosome 9. Nature. 2004;**429**:369-374

[72] Lozano G, Francisco-Velilla R, Martinez-Salas E. Deconstructing internal ribosome entry site elements: An update of structural motifs and functional divergences. Open Biology. 2018;**8**:180155

[73] Du X, et al. Second cistron in *CACNA1A* gene encodes a transcription factor mediating cerebellar development and SCA6. Cell. 2013;**154**: 118-133

[74] Xue S, Tian S, Fujii K, Kladwang W, Das R, Barma M. RNA regulons in *Hox* 5' UTRs confer ribosome specificity to gene regulation. Nature. 2015;**517**:33-38

[75] de Mateo S, Castillo J, Estanyol JM, Ballescà JL, Oliva R. Proteomic characterization of the human sperm nucleus. Proteomics. 2011;**11**:2714-2726

[76] Waddington CH. Genetic assimilation of the bithorax phenotype. Evolution. 1956;**10**:1-13

[77] Otaki JM, Hiyama A, Iwata M, Kudo T. Phenotypic plasticity in the range-margin population of the lycaenid butterfly *Zizeeria maha*. BMC Evolutionary Biology. 2010;**10**:252

[78] Hiyama A, Taira W, Otaki JM. Color-pattern evolution in response to environmental stress in butterflies. Frontiers in Genetics. 2012;**3**:15

[79] Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;**478**:343-348

[80] Harcourt AH, Harvey PH, Larson SG, Short RV. Testis weight, body weight and breeding system in primates. Nature. 1981;**293**:55-57

[81] Yamada Y, et al. Identification of nine genes as novel susceptibility loci for early-onset ischemic stroke, intracerebral hemorrhage, or subarachnoid hemorrhage. Biomedical Reports. 2018;**9**:8-20