

Livebearing or egg-laying mammals: 27 decisive nucleotides of FAM168

Subrata Pramanik¹, Arne Kutzner², Klaus Heese^{1,*}

¹ Graduate School of Biomedical Science and Engineering, Hanyang University, Seoul, Korea;

² Department of Information Systems, College of Engineering, Hanyang University, Seoul, Korea.

Summary

In the present study, we determine comprehensive molecular phylogenetic relationships of the novel myelin-associated neurite-outgrowth inhibitor (MANI) gene across the entire eukaryotic lineage. Combined computational genomic and proteomic sequence analyses revealed MANI as one of the two members of the novel family with sequence similarity 168 member (FAM168) genes, consisting of FAM168A and FAM168B, having distinct genetic differences that illustrate diversification in its biological function and genetic taxonomy across the phylogenetic tree. Phylogenetic analyses based on coding sequences of these FAM168 genes revealed that they are paralogs and that the earliest emergence of these genes occurred in jawed vertebrates such as *Callorhinchus milii*. Surprisingly, these two genes are absent in other chordates that have a notochord at some stage in their lives, such as branchiostoma and tunicates. In the context of phylogenetic relationships among eukaryotic species, our results demonstrate the presence of FAM168 orthologs in vertebrates ranging from *Callorhinchus milii* to *Homo sapiens*, displaying distinct taxonomic clusters, comprised of fish, amphibians, reptiles, birds, and mammals. Analyses of individual FAM168 exons in our sample provide new insights into the molecular relationships between FAM168A and FAM168B (MANI) on the one hand and livebearing and egg-laying mammals on the other hand, demonstrating that a distinctive intermediate exon 4, comprised of 27 nucleotides, appears suddenly only in FAM168A and there in the livebearing mammals only but is absent from all other species including the egg-laying mammals.

Keywords: Central nervous system, genomics, evolution, eukaryotes, FAM168, myelin, neuron, phylogenetic, orthologs

1. Introduction

Understanding phylogenetic gene distributions is one of the most challenging aspects of modern genomics-supported taxonomy (1-3). A primary goal is to understand the molecular basis of phylogenetics, which is necessary to determine the origins of species (4,5). In this context, species specification is crucial for the determination of a potential evolutionary process (3,6). Large-scale comparative genomics analyses

have revealed that gene duplication and mutations are pervasive sources of genetic changes that underlie phenotypic diversity among species (7,8). Despite a longstanding interest in the genetic basis of speciation, little is known about genetic changes in the human lineage or their implications in human evolution theory (9,10).

Recent progress in sequencing technologies has provided unprecedented opportunities for exploring genetic differences between primitive and derived species (11). The increased availability of new sequence data, e.g., DNA sequences, mRNA expression, and proteins, may not directly provide fundamental knowledge about speciation or interspecies relationships (12). However, comparative analyses of sequence data across the phylogenetic tree can provide insights into detailed speciation pathways (13-16).

Recently, our group identified and characterized the

Released online in J-STAGE as advance publication April 3, 2017.

*Address correspondence to:

Dr. Klaus Heese, Graduate School of Biomedical Science and Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Republic of Korea.

E-mail: klaus@hanyang.ac.kr

novel human neuronal protein, family with sequence similarity 168 member B (FAM168B also known as myelin-associated neurite-outgrowth inhibitor (MANI)), which is a member of the FAM168 family (17,18). FAM168B is localized to neuronal cell membranes and has potential for inhibition of neurite-outgrowth and axonal guidance in the central nervous system (CNS). Our findings also suggested that FAM168B plays an important role in neuronal differentiation of neural stem cells (NSCs) into catecholaminergic neurons (17,18). Other studies have recently characterized the human FAM168A gene (also known as tongue cancer resistance-associated protein 1 (TCRP1)) in oral squamous cell carcinoma (OSCC) cells (19-21). These studies demonstrated that FAM168A mediates specific resistance to cisplatin in Tca8113 cells by reducing cisplatin-induced apoptosis (20,22). Available data show that FAM168A and FAM168B have distinct physiological functions, even though they belong to the same gene family and exhibit very high gene homology. Accordingly, we performed comparative genomic and proteomic sequence analyses to explore further potential functional implications of FAM168. Phylogenetic analyses of this gene family across the entire eukaryotic tree of life revealed the phylogenetic origins and taxonomic relationships of and among the species carrying these genes, demonstrating that a distinctive intermediate exon, comprising 27 nucleotides (nts) only, appears in FAM168A and defines livebearing mammals.

2. Materials and Methods

2.1. Materials

The human chromosome (chr) dataset used for the genomic, proteomic, and phylogenetic analyses in the present study was collected from the public database of the National Center for Biotechnology Information (NCBI). The dataset for FAM168A, which is located on chr 11, has access number NC_000011.10 (chr 11, GRCh38), and the dataset for FAM168B, which is located on chr 2, has access number NC_000002.12 (chr 2, GRCh38). Details of FAM168 gene IDs, mRNA, and protein sequence sources used in this study are provided as Supplementary materials (Figure S1, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>).

2.2. Data analysis and sequence alignments

In the present study, we used several genome viewer tools, including Integrative Genomics Viewer (IGV) (23,24), NCBI Map Viewer (25-28), the UCSC genome browser (29,30), and ClinVar (31) for the visualization and analysis of genomic data. For local alignments, a whole human genome analysis of FAM168 gene family homology search was performed using NCBI's BLAST

program, which finds regions of local similarity between sequences (32,33). For global multi-alignments, the retrieved mRNA and protein sequences were aligned using multiple alignment tools, including Clustal Omega (34) and MUSCLE (MULTiple Sequence Comparison by Log-Expectation) (35). All standard parameters were unchanged unless stated otherwise. The alignment was optimized manually according to previous knowledge of exons and coding sequences (CDSs) based on visualizations using genome viewing tools such as IGV and NCBI Map Viewer.

2.3. FAM168 analysis in the genus *Homo*

In order to analyse FAM168 genes in *Homo*, including *Homo Neandertalensis* and Denisovan samples, FASTQ reads provided as BAM files by the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany (eva.mpg.de), were extracted and realigned locally to the GRCh38 assembly of *H. sapiens* with the Bowtie2 aligner. All standard parameters were unchanged unless stated otherwise (15,36) (Supplementary methods).

2.4. FAM168 in phylogenetic analysis

We developed a special tailored C++11 application with an embedded Burrows-Wheeler Aligner as the central component (15,37) (Supplementary methods). Using the integrated aligner, our application allowed for the tracing of individual sequences within a taxonomic context. We relied on a taxonomy offered by NCBI that can be reconstructed based on a foundation of publically available database dumps (38,39). The generated phylogenetic trees were visualized and analyzed using Archaeopteryx (40). Additionally, we retrieved the genomes of species within the taxonomy from assemblies offered by NCBI (28,38,39), and all data retrievals were conducted automatically by evaluating database-information available from NCBI (15,41,42).

3. Results

3.1. Chr loci of FAM168 genes

Whole human genome analysis of the FAM168 gene family using a small nt homology search in BLAST revealed that there are two members of this gene family, FAM168A and FAM168B. Both FAM168A and B are transcribed on the reverse strand (Figure 1). Comparisons of all exons of FAM168A and FAM168B showed that the two gene members are paralogous in humans, and that two intermediate exons are missing in the longest isoform of member B (NM_001009993.3) having 5449 nts in its mRNA and 195 amino acids (aa) in its protein with respect to the longest isoform of member A (NM_001286050.1) and having 7305 nts in its mRNA and 244 aa in its protein (Figures

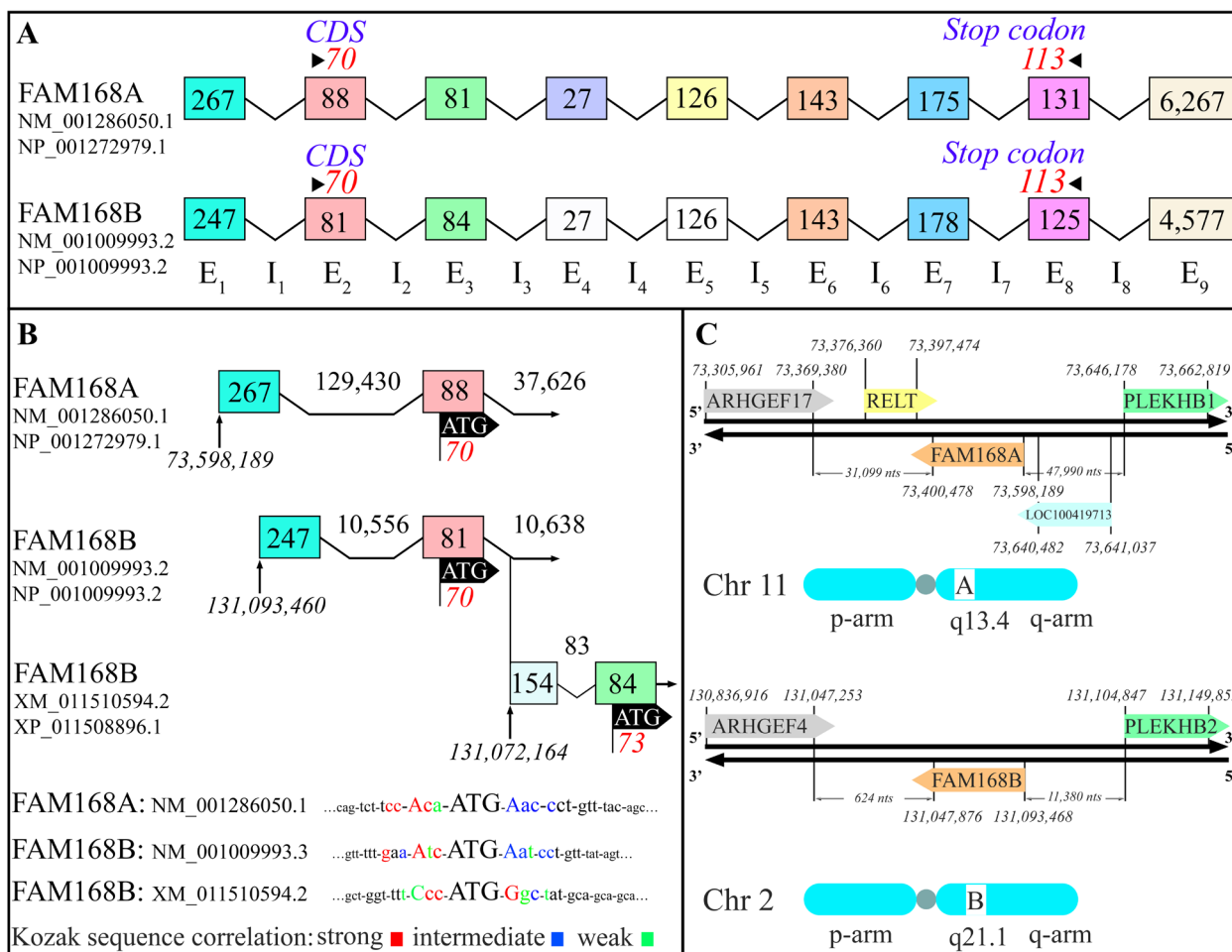


Figure 1. Gene structures of FAM168A and FAM168B. (A) Homology between FAM168A and FAM168B. The CDSs start from exon 2 (E2) and stop at E8 for both genes. Filled boxes of different colors indicate homologous sequences present in both genes, whereas the empty boxes of FAM168B (E4, E5) indicate that these two exons are absent in FAM168B but present in FAM168A. For consistency, exon and intron counts for FAM168B were made with respect to FAM168A. The intermediate two exons and two introns of FAM168B (E4, E5, I4, and I5) are missing with respect to FAM168A in the livebearing mammals. (B) Transcripts of start codons and Kozak consensus sequence of FAM168s. A newly predicted short transcript of FAM168B (168 aa) has a relatively strong Kozak consensus sequence, as do the longer transcripts. This short version remains to be confirmed by further experiments. (C) Comparative human genomic loci of FAM168A and FAM168B. FAM168A is located in the q arm of chr 11, whereas FAM168B is located in the q arm of chr 2. Neighboring genes of FAM168s belong to members of the ARHGEF and PLEKHB families, respectively. An additional gene, RELT is present between FAM168A and ARHGEF17, but not between FAM168B and ARHGEF4, thus suggesting a possible function of FAM168A in the immune system.

1 and S2, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). According to isoform analysis, all isoforms of FAM168A have the same start and stop codons, whereas variation was observed in the intermediate exons for both validated and predicted isoforms (Figures 1 and S2, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Similarly, all validated and predicted isoforms of FAM168B have the same start and stop codons, except the predicted short isoforms XM_017003328.1 and XM_011510594.2 (Figures 1 and S3, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>).

3.2. FAM168 transcripts

According to NCBI's GRCh38.p7 gene assembly,

FAM168A has validated short mRNA transcripts of 7305 nts (244 aa), 7278 nts (235 aa), and 6960 nts (129 aa) as well as predicted (using NCBI's Gnomon software (43,44)) short isoforms of 6571 nts (193 aa) and 6406 nts (138 aa) using the same ATG-start codon with an alternative splicing pattern (Figures 1B, S2, and S3, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). This observation is similar to the earlier GRCh38.p2 gene assembly, which also predicted a short FAM168A transcript (6776 nts, 193 aa) using the same start codon as the long transcript (7305 nts, 244 aa).

For FAM168B, GRCh38.p7 gene assembly analysis showed multiple transcripts with CDSs for 195 aa (e.g., 5449 nts, 5590 nts, 5884 nts, 5331 nts, 5338 nts, 5258 nts, 5207 nts, and 5927 nts) using the same ATG-start codon with an alternative splicing pattern

Table 1. Lengths and GC and AT contents of FAM168A (NM_001286050.1) and FAM168B (NM_001009993.3)

Exon	# nts in FAM168A			# nts in FAM168B		
	length	GC%	AT%	length	GC%	AT%
1	267	73.40	26.60	247	78.94	21.06
2	88	55.68	44.32	81	39.50	60.50
3	81	58.02	41.98	84	51.19	48.81
4	27	40.74	59.26	–	–	–
5	126	57.14	42.86	–	–	–
6	143	56.64	43.36	143	62.23	37.77
7	175	61.14	38.86	178	62.92	37.08
8	131	60.30	39.70	125	63.20	36.80
9	6,267	48.69	51.31	4,577	45.48	54.52
5'UTR	285	71.92	28.08	258	75.98	24.02
CDS	735	57.95	42.05	588	59.18	40.82
3'UTR	6,285	48.73	51.27	4,589	45.47	54.53

(Figure S3, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Additionally, the predicted short transcripts of 168 aa (e.g., 5541 nts and 5250 nts) were observed using the fourth in-frame ATG-start codon of its full-length transcript (Figure S3, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Importantly, the start codon of the short prediction of 168 aa also has a relatively strong Kozak sequence compared with the start codon used by the longer transcript (Figure 1B).

3.3. FAM168 CDSs and protein sequence comparisons

A comparative CDS analysis of the longest isoform of FAM168A (NM_001286050.1, NP_001272979.1) and FAM168B (NM_001009993.3, NP_001009993.2) showed that two intermediate exons comprised of 27 and 126 nts are missing in FAM168B compared to FAM168A (Figures 1A, S2, and S3, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Protein sequence comparison of the two FAM168 proteins encoded by chr2 and chr11 also revealed significant differences at the protein level based on nt differences in the CDSs (Figure 1 and Table 1).

3.4. 5'- untranslated regions (5'-UTRs) and 3'-UTRs of FAM168

According to UTRs analysis, 5'-UTRs are significantly shorter than 3'-UTRs for both FAM168A and FAM168B (Table 1). A comparison of the 5'-UTR of FAM168A (285 nts of NM_001286050.1) with the 5'-UTR of FAM168B (258 nts of NM_001009993.3) indicated that they are not significantly homologous (Figures S4 and S5, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Likewise, 3'-UTRs of FAM168A (6,285 nts of NM_001286050.1) and FAM168B (4,589 nts of NM_001009993.3) were also not significantly homologous (Figures S4 and S5, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>).

3.5. GC and AT content of FAM168

A comparative analysis of GC and AT contents of FAM168A and FAM168B showed that their respective percentage contents have high similarity except for two intermediate exons of FAM168A containing 27 and 126 nts, which are missing in FAM168B (Table 1). In the 5'-UTR regions, GC% (71.92% for FAM168A and 75.98% for FAM168B) are significantly higher than AT% (28.08% for FAM168A and 24.02% for FAM168B) for both FAM168A and FAM168B. In the CDSs, GC and AT contents are almost identical for FAM168A and FAM168B, respectively. However, GC content is higher than AT content in both FAM168A and FAM168B. Interestingly, exon 2 of FAM168B, containing the start codon, has a relatively low GC content of about 39.5%, while the average GC content of a gene is usually in the range of 50%-60% (45). On the other hand, GC contents (48.73% for FAM168A and 45.47% for FAM168B) are lower than AT (51.27% for FAM168A and 54.53% for FAM168B) contents in the 3'-UTRs (Table 1).

3.6. Closest neighboring genes of FAM168

Analyses of neighboring loci of human FAM168 revealed complex relationships of FAM168A and FAM168B with their neighbor genes. We observed that the common genes pleckstrin homology domain containing B1 (PLEKHB1) and PLEKHB2 reside upstream of FAM168A and FAM168B, respectively (Figure 1C). The intermediate length between FAM168A and PLEKHB1 (47,990 nts) differs from that between FAM168B and PLEKHB2 (11,380 nts). Moreover, based on NCBI's GRCh38 gene assembly, we observed that the cutaneous T-cell lymphoma-associated antigen (CTAGE) family member 5, a pseudogene (LOC100419713), is situated between FAM168A and PLEKHB1, whereas no other genes were observed between FAM168B and PLEKHB2. A similar phenomenon is observed in the downstream analysis of FAM168A and FAM168B (Figure 1C). The common genes Rho guanine nucleotide exchange

factor 17 (ARHGEF17) and ARHGEF4 (members of the ARHGEF gene family) reside downstream of both FAM168A and FAM168B. However, the receptor expressed in lymphoid tissues (RELT) gene is located only between FAM168A and ARHGEF17, not between FAM168B and ARHGEF4, thus indicating a possible function of FAM168A in the immune system. The intermediate length between FAM168A and ARHGEF17 (31,099 nts) significantly differs from that between FAM168B and ARHGEF4 (624 nts).

3.7. FAM168 in *Homo*

In search of the closest hominin relative of *H. sapiens*, recent discoveries of genomic data obtained by sequencing ancient DNA from Neandertal and Denisovan fossils might enable us to answer longstanding questions about the relationships between archaic and modern humans (46,47). In the present study, we analyzed the FAM168 gene family of *H. sapiens* and compared it with data from Neandertals and Denisovans. FAM168A and FAM168B are found in both Neandertals and Denisovans, although with different mutation patterns (see Tables S1 and S2 for details, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Exon analyses among groups indicate that all exons of FAM168A are identical among the three *Homo* genomes except E7 and E9, whereas all introns of FAM168A display a number of mutations (Figure S6, Tables S1 and S2, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). *H. sapiens* differs by one nt from Neandertals and Denisovans in E7 of FAM168A. However, variations in E9 of FAM168A remain elusive, where three nts of *H. sapiens* differ from Neandertals and Denisovans, six nts of Neandertals differ from *H. sapiens* and Denisovans, and four nts of Denisovans differ from *H. sapiens* and Neandertals (see Tables S1 and S2 for details, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). In the case of FAM168B, all exons are identical among the three *Homo* genomes except E9. For consistency, exon and intron counts for FAM168B were made with respect to FAM168A. The intermediate two exons and two introns of FAM168B (E4, E5, I4, and I5) are missing with respect to FAM168A in the genus *Homo*. The variations in E9 of FAM168B include *i*) one nt difference in *H. sapiens* compared with Neandertals and Denisovans, *ii*) Neandertals differ by three nts from Denisovans and *H. sapiens*, and *iii*) Denisovans differ by one nt from *H. sapiens* and Neandertals. Intron analysis of FAM168B showed a number of mutations in all introns except I6 and I8 (see Figure S6, Tables S1 and S2 for details, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Although Neandertals and Denisovans are considered the closest hominin relatives of *H. sapiens* (48), our genomic comparison does not indicate which archaic human is more closely

related to *H. sapiens*. Nevertheless, growing evidence suggest that Neandertals are more closely related to *H. sapiens* than are Denisovans (15,49,50).

3.8. FAM168 in phylogenetic analyses

We conducted a comparative genomic analysis to explore the phylogenetic distribution of the FAM168 gene family among the eukaryotes. The phylogenetic relationships of FAM168 gene families among different species are displayed in the form of taxonomic clusters, or dendrograms, through CDS sequence alignments (15,41,42) (Figures S7 and S8, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Aligned sequences appear as taxonomic cluster blocks of high correlation along the diagonal axis and correspond to taxonomic groups (Figure 2). According to the phylogenetic tree, both FAM168A and FAM168B formed five distinct ortholog clusters of fish, amphibians, reptiles, birds, and mammals. Interspecies correlations showed that both genes have similar phylogenetic patterns from lower species to higher species. The sequence similarity matrix in the dendrograms is colored as a heat map (Figure 2). Within the mammals, FAM168A shows higher interspecies correlation than FAM168B. Interestingly, among the birds, *Passeriformes* showed a distanced sub-cluster ortholog for both genes with higher sequence similarity (Figure 2). Within the reptiles, two distinct sub-clusters are formed by the homologs of FAM168 genes, one including turtles and crocodylians and another within lepidosaurs. Unlike the mammals, available genomic data reflects that FAM168B has higher interspecies correlation in reptiles than does FAM168A. In the context of phylogenetic relationships between fish and reptiles, the intermediate class of amphibians (*e.g.*, tropical clawed frog) also contains both genes (51). The earliest apparent emergence of both FAM168A and FAM168B is observed in the jawed vertebrates, represented by *Callorhynchus milii* (elephant shark) (gene ID: 103177153 for FAM168A and gene ID: 103174665 for FAM168B) (25,52) (Figures 2, S7 and S8, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Analysis of the available molecular phylogenetic data suggests that the FAM168 gene pair emerged in vertebrates with a notochord and neural tube (53).

Our comparative genomic and proteomic analyses showed that two intermediate exons comprised of 27 and 126 nts (E4 and E5), respectively, are missing in FAM168B with respect to FAM168A in humans (Figures 1A, S4 and S8, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Further phylogenetic analysis of the individual exons in the eukaryotic lineage revealed that exon 4 of FAM168A, comprised of 27 nts (E4, translated into nine aa: EFQFLHSAY), is present in the livebearing marsupial *Monodelphis domestica*

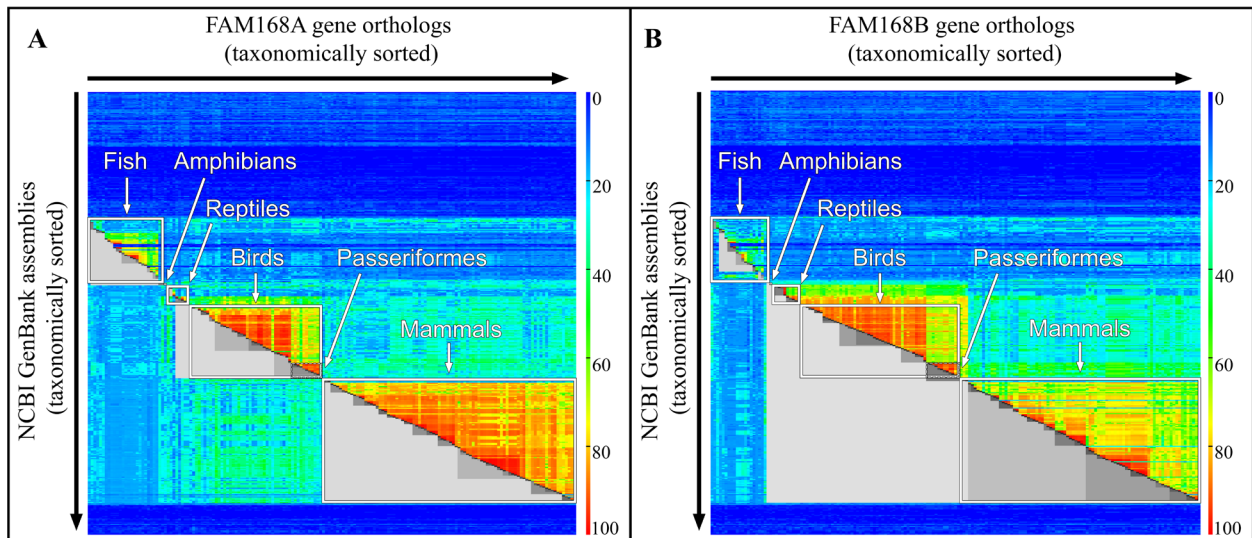


Figure 2. Cluster representation of phylogenetic analysis of FAM168A and FAM168B in the eukaryotic lineage, based on the NCBI GenBank. (A) Orthologs of FAM168A displayed discrete taxonomic cluster groups with higher correlation along the diagonal (red = high homology, blue = low homology). Taxonomic cluster blocks along the diagonal correspond to (i) fish, (ii) amphibians, (iii) reptiles, (iv) birds, and (v) mammals. **(B)** Orthologs of FAM168B displayed a similar pattern of sequence similarity with discrete taxonomic cluster groups with higher correlation along the diagonal. Interestingly, FAM168B showed cross-clustering or cluster overlapping between reptiles and birds as well as between birds and mammals. *Passeriformes* showed a sub-cluster within the birds. The gray triangles below the diagonal display clusters obtained when using a threshold of 60 as described previously (54).

(gray short-tailed opossum) but is absent in the egg-laying mammal *Ornithorhynchus anatinus* (platypus) and in all other species in the analysis, including birds (e.g., *Ficedula albicollis*), reptiles (e.g., *Alligator mississippiensis*), amphibians (e.g., *Xenopus tropicalis*), and fish (e.g., *Danio rerio* and *C. milii*) (Figures 3, S9, and S10, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). In contrast, exon 5 of FAM168A, comprised of 126 nts (E5, translated into 42 aa in *H. sapiens*), is conserved in the mammals as well as in birds (e.g., *Ficedula albicollis*), reptiles (e.g., *Alligator mississippiensis*), amphibians (e.g., *Xenopus tropicalis*), and fish (e.g., *Danio rerio* and *C. milii*) (Figures 3, S9, and S10, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Protein sequence comparisons indicate that significant differences exist between these two proteins, FAM168A and FAM168B (Figures 3 and S4, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). These observations suggest that higher mutations occurred in regions located closer to the N-terminus for FAM168 proteins during possible evolutionary and speciation events.

4. Discussion

4.1. Genomics and proteomics analyses of FAM168

We provide for the first time a comprehensive genomics and proteomics feature overview of the FAM168 gene family. We found that the human FAM168A and FAM168B paralogs are located on chr 11 and chr 2, respectively, and show significant sequence differences

(Figures 1 and 3). In particular, protein sequence comparisons showed that most variations appear toward the N-terminus of FAM168 proteins. Analyses of the individual exons suggest that deletion and/or insertion occurred during gene duplication events, leading to the emergence of new genes within the FAM168 gene family (55-58) (Figures 3, S9, and S10, <http://www.biosciencetrends.com/action/getSupplementalData.php?ID=9>). Experimental data suggest that FAM168B (MANI) plays a role in neuro-axonal guidance and neuronal differentiation in the CNS (17,18), whereas FAM168A functions in chemoresistance by reducing cisplatin-mediated apoptosis (20,22). Thus, genetic differences between the FAM168A and FAM168B may explain why these two genes have distinct functions (2,56,59).

4.2. Phylogenetic analysis of FAM168

This is the first detailed phylogenetic analysis of the FAM168 gene family across the eukaryotic lineage. Our phylogenetic analyses, based on the CDSs of the two FAM168 genes (A and B), outline deep relationships among eukaryotes. The earliest emergence of the FAM168 genes may have occurred in the jawed vertebrates, represented by *C. milii*, which is surprising as these two genes are absent in the other main chordate sub-group (non-vertebrate chordates) that also have a notochord at some stage in their lives, for example, branchiostoma (e.g., *Branchiostoma belcheri* and *Branchiostoma floridae*) and tunicates (e.g., *Botryllus schlosseri*, *Ciona intestinalis*, *Oikopleura dioica*, and *Ciona savignyi*) (52,60,61). Accordingly, FAM168A

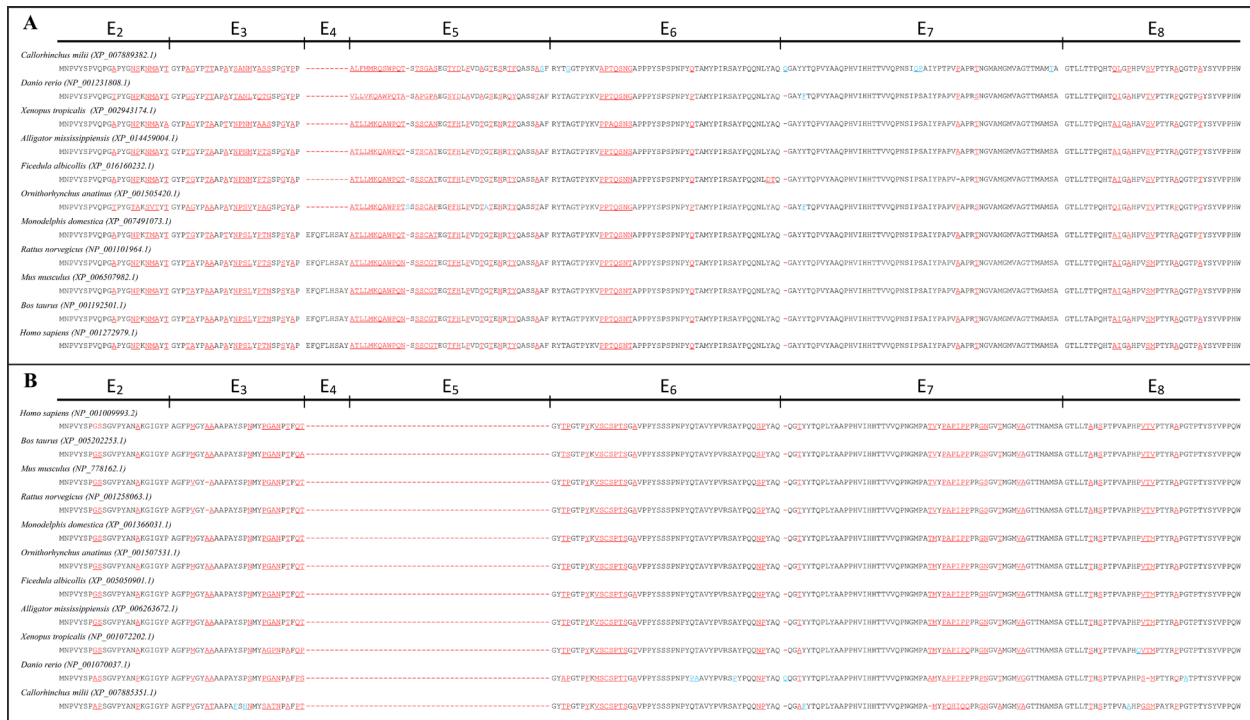


Figure 3. Comparative protein sequence analysis of FAM168A and FAM168B. FAM168A and FAM168B proteins were observed earliest in the jawed vertebrates, represented here by *C. milii*. (A) Protein sequence analysis of FAM168A for a few species, from *C. milii* to *H. sapiens*. Analyses showed that exon 4 of FAM168A, comprised of 27 nts (E4, translated into nine aa: EFQFLHSAY), is present in livebearing marsupials, represented by *M. domestica*, but is absent in egg-laying mammals such as *O. anatinus* and in all other groups including fish, represented by *C. milii*. Exon 5 of FAM168A, comprised of 126 nts (E5, translated into 42 aa of *H. sapiens*), is conserved across all groups. (B) Protein sequence analysis of FAM168B is shown for a few representative species across the phylogenetic tree, from *C. milii* to *H. sapiens*. All exons (E1-E9) are presented in Figures 1A and S4.

and FAM168B seem to be distinctive features of vertebrates (Figure 4).

Individual cluster analysis demonstrates that both genes are highly conserved within each cluster of species; however, FAM168B showed cross-clustering for birds with reptiles and birds with mammals (Figure 2). Although both FAM168 genes show a similar phylogenetic distribution in the Callorhinchidae, a derived mutation pattern is observed in *H. sapiens*. However, we also identified a few species within the fish, amphibians, reptiles, birds, and mammals (e.g., *Melanochromis auratus*, *Nanorana parkeri*, *Apalone spinifera*, *Phoenicopterus ruber*, *Megaderma lyra*, *Manis pentadactyla*, and *Apodemus sylvaticus*) without either FAM168A or FAM168B. We cannot rule out that we failed to detect these genes in some species because genome data for some species remain incomplete or at the scaffold level only.

The exon–intron architecture is a longstanding question in evolutionary genomics (62-65). Changes in the splicing of exons and introns are a major driving force in proteomic diversification and the generation of new gene functions (64,66,67). In our molecular phylogenetic analysis, we observed that FAM168A contains additional exons E4 and E5 (with respect to FAM168B) across all vertebrate species (Figures 3, S9, and S10, <http://www.biosciencetrends.com/>

action/getSupplementalData.php?ID=9). However, while exon E5 is present in FAM168A across all vertebrates, the jawed vertebrates, represented by *C. milii*, do not contain E4, whereas *H. sapiens* contains E4 in FAM168A (Figures 3 and 4). Thus, we sought to identify the phylogenetic origin of this distinctive 27 nts-containing exon E4 in FAM168A. Surprisingly, we observed this exon in the livebearing mammal *M. domestica*, but not in the egg-laying mammal *O. anatinus*. Thus, this intermediate exon E4 seems to be a distinctive feature of livebearing mammals. Moreover, insertion of this new exon may have led to proteomic diversification by generating new gene functionality in FAM168A, the development of a higher immune system, which is essential for maturation in livebearing mammals (56,67,68) (Figure 4). Although the immune system is relatively undeveloped at birth and is developed during a lifetime of exposure to multiple foreign challenges (the so-called adaptive immune system), the development of the immune system, in particular innate immunity, starts early in fetal life for livebearing mammals (69-71). Considering this finding and our previous discoveries, the FAM168 gene family may contain crucial genes involved in the higher immune system and brain functions that are essential for mammals giving birth to live young (15,17,18,68).

Concluding, our results reflect a comprehensive

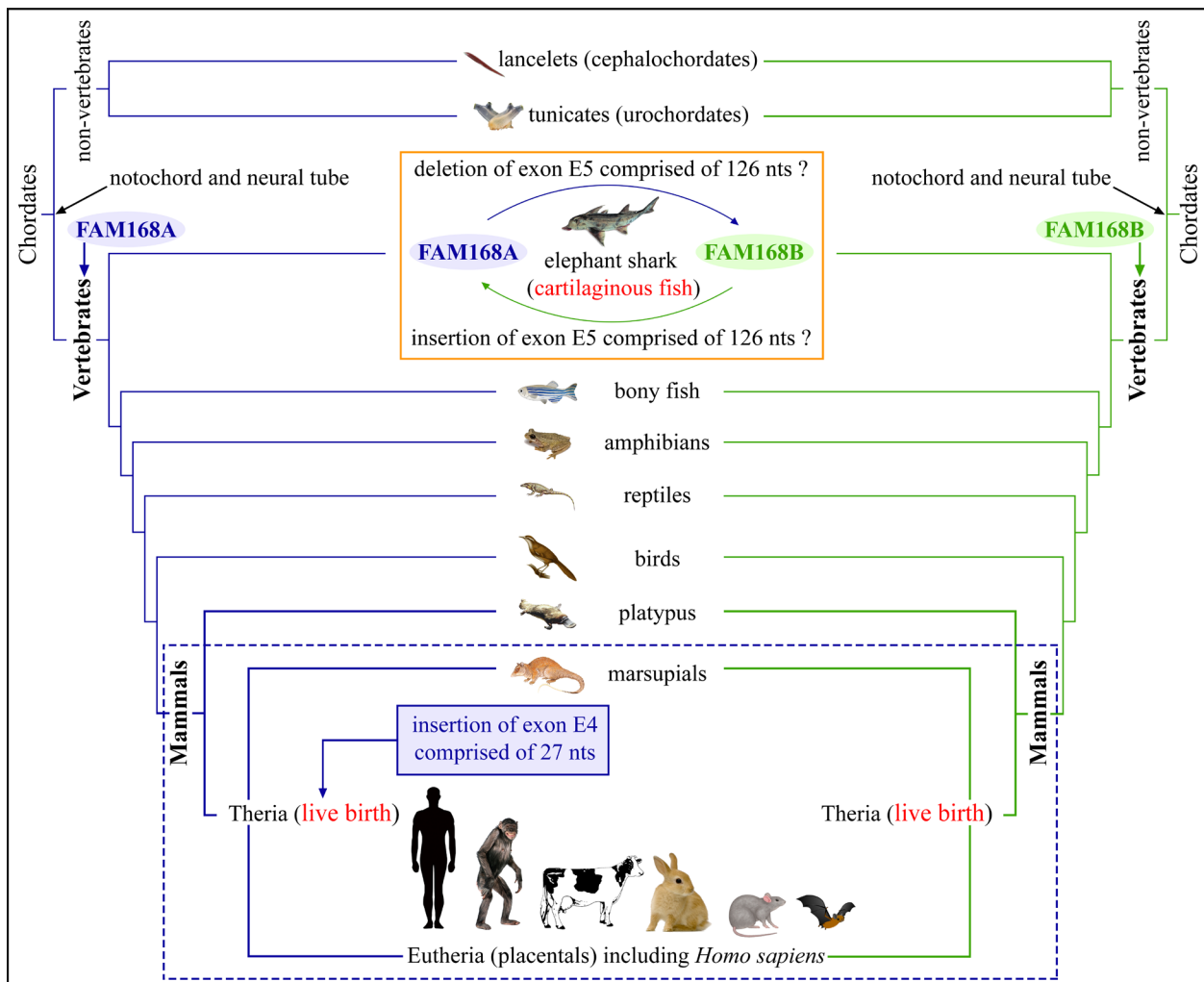


Figure 4. Phylogenetic overview of FAM168A and FAM168B from elephant shark to human in the eukaryotic lineage. Both FAM168A and FAM168B appear only in the vertebrates. Interestingly, only animals that give birth to live young have the distinctive exon E4, comprising 27 nts, in FAM168A, implying a functional significance in terms of higher immune system and brain function development.

picture of the genomic, proteomic, and phylogenetic features of the FAM168 gene family. We identified significant differences between FAM168A and FAM168B, in particular the incorporation of exon E5 into FAM168A in lower vertebrate species, such as *C. milii*, and the sudden incorporation of the distinctive exon E4 into FAM168A in higher vertebrate species (*i.e.* mammals) that give birth to live young. These patterns may illustrate functional diversification and species-dependent functional specification across the phylogenetic tree of the eukaryotic lineage (62,64,65). The phylogenetic distributions of FAM168A and FAM168B from *C. milii* to *H. sapiens* comprise distinct taxonomic clusters of species, thus indicating that morphology-based analyses remain insufficient to accurately define the relationships among species (2). Genomic analysis across a large sample of species may allow the identification of interspecies relationships and phylogenetic hot spots of distinctive gene origins (8,42). Future experimental studies investigating the regulation of gene expression and function of FAM168

along with large-scale gene datasets from additional diverse lineages using an even more global perspective may provide further insights into the functional significance of the two FAM168 genes across the entire phylogenetic tree of life, particularly in terms of specific higher immune and brain functions.

Acknowledgements

This study was supported by Hanyang University. We thank the Max Planck Institute, Evolutionary Anthropology, Leipzig, Germany (eva.mpg.de) for providing the Neandertal and Denisova genome data. We thank Mr. Markus Schmidt for technical assistance.

References

1. Carroll SB. Genetics and the making of *Homo sapiens*. Nature. 2003; 422:849-857.
2. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005; 39:309-338.

3. Seehausen O, Butlin RK, Keller I, *et al.* Genomics and the origin of species. *Nat Rev Genet.* 2014; 15:176-192.
4. Necșulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet.* 2014; 15:734-748.
5. Chan YF. Hearing echoes. *Heredity.* 2012; 108:471-472.
6. Noor MA, Feder JL. Speciation genetics: Evolving approaches. *Nat Rev Genet.* 2006; 7:851-861.
7. Takemura M, Yokobori S, Ogata H. Evolution of Eukaryotic DNA Polymerases *via* Interaction Between Cells and Large DNA Viruses. *J Mol Evol.* 2015; 81:24-33.
8. Burki F. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb Perspect Biol.* 2014; 6:a016147.
9. Gilbert SL, Dobyns WB, Lahn BT. Genetic links between brain development and brain evolution. *Nat Rev Genet.* 2005; 6:581-590.
10. Burgess DJ. Evolutionary genetics: Haunted by the past-modern consequences of Neanderthal DNA. *Nat Rev Genet.* 2016; 17:191.
11. Paabo S. The diverse origins of the human gene pool. *Nat Rev Genet.* 2015; 16:313-314.
12. Noguchi F, Tanifuji G, Brown MW, Fujikura K, Takishita K. Complex evolution of two types of cardiolipin synthase in the eukaryotic lineage stramenopiles. *Mol Phylogenet Evol.* 2016; 101:133-141.
13. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A.* 2003; 100:3351-3356.
14. Werner-Washburne M, Wylie B, Boyack K, Fuge E, Galbraith J, Weber J, Davidson G. Comparative analysis of multiple genome-scale data sets. *Genome Res.* 2002; 12:1564-1573.
15. Kutzner A, Pramanik S, Kim PS, Heese K. All-or-(N) One - an epistemological characterization of the human tumorigenic neuronal paralogous *FAM72* gene loci. *Genomics.* 2015; 106:278-285.
16. Wei K, Li Y, Chen H, Zhang Q. Genomic Surveillance Elucidates HCV 1a Phylodynamics and Molecular Evolution. *Evol Biol.* 2016; 43:380-391.
17. Mishra M, Lee S, Lin MK, Yamashita T, Heese K. Characterizing the neurite outgrowth inhibitory effect of Mani. *FEBS Lett.* 2012; 586:3018-3023.
18. Mishra M, Akatsu H, Heese K. The novel protein MANI modulates neurogenesis and neurite-cone growth. *J Cell Mol Med.* 2011; 15:1713-1725.
19. Gu Y, Fan S, Liu B, Zheng G, Yu Y, Ouyang Y, He Z. TCRP1 promotes radioresistance of oral squamous cell carcinoma cells *via* Akt signal pathway. *Mol Cell Biochem.* 2011; 357:107-113.
20. Gu Y, Fan S, Xiong Y, Peng B, Zheng G, Yu Y, Ouyang Y, He Z. Cloning and functional characterization of *TCRP1*, a novel gene mediating resistance to cisplatin in an oral squamous cell carcinoma cell line. *FEBS Lett.* 2011; 585:881-887.
21. Peng B, Yi S, Gu Y, Zheng G, He Z. Purification and biochemical characterization of a novel protein-tongue cancer chemotherapy resistance-associated protein1 (TCRP1). *Protein Expr Purif.* 2012; 82:360-367.
22. Liu X, Wang C, Gu Y, Zhang Z, Zheng G, He Z. TCRP1 contributes to cisplatin resistance by preventing Pol β degradation in lung cancer cells. *Mol Cell Biochem.* 2015; 398:175-183.
23. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol.* 2011; 29:24-26.
24. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform.* 2013; 14:178-192.
25. Wolfsberg TG. Using the NCBI Map Viewer to browse genomic sequence data. *Curr Protoc Hum Genet.* 2011; Chapter 18:Unit18 15.
26. Tatusova T. Genomic databases and resources at the National Center for Biotechnology Information. *Methods Mol Biol.* 2010; 609:17-44.
27. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic acids research.* 2016; 44:D7-19.
28. Sayers EW, Barrett T, Benson DA, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2010; 38:D5-16.
29. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. *Brief Bioinform.* 2013; 14:144-161.
30. Sanborn JZ, Benz SC, Craft B, *et al.* The UCSC Cancer Genomics Browser: Update 2011. *Nucleic Acids Res.* 2011; 39:D951-959.
31. Harrison SM, Riggs ER, Maglott DR, Lee JM, Azzariti DR, Niehaus A, Ramos EM, Martin CL, Landrum MJ, Rehm HL. Using ClinVar as a Resource to Support Variant Interpretation. *Curr Protoc Hum Genet.* 2016; 89:8.16.11-18.16.23.
32. Mount DW. Using the basic local alignment search tool (BLAST). *CSH Protoc.* 2007; 2007:pdb top17.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; 215:403-410.
34. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol.* 2014; 1079:105-116.
35. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004; 32:1792-1797.
36. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012; 9:357-359.
37. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010; 26:589-595.
38. Sayers EW, Barrett T, Benson DA, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2009; 37:D5-15.
39. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2009; 37:D26-31.
40. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics.* 2009; 10:356.
41. Bonder MJ, Abeln S, Zaura E, Brandt BW. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics.* 2012; 28:2891-2897.
42. Stoeckle MY, Coffran C. TreeParser-aided Klee diagrams display taxonomic clusters in DNA barcode and nuclear gene datasets. *Sci Rep.* 2013; 3:2635.
43. Souvorov A, Kapustin Y, Kiryutin B, Chetvernin V, Tatusova T, Lipman D. Gnomon-NCBI eukaryotic gene prediction tool. *National Center for Biotechnology Information.* 2010;1-24.

44. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* 2012; 40:D130-135.
45. Zhang L, Kasif S, Cantor CR, Broude NE. GC/AT-content spikes as genomic punctuation marks. *Proc Natl Acad Sci U S A.* 2004; 101:16855-16860.
46. Slatkin M, Racimo F. Ancient DNA and human history. *Proc Natl Acad Sci U S A.* 2016; 113:6380-6387.
47. Vernot B, Tucci S, Kelso J, *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science.* 2016; 352:235-239.
48. Prufer K, Racimo F, Patterson N, *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature.* 2014; 505:43-49.
49. Sawyer S, Renaud G, Viola B, Hublin JJ, Gansauge MT, Shunkov MV, Derevianko AP, Prufer K, Kelso J, Paabo S. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proc Natl Acad Sci U S A.* 2015; 112:15696-15700.
50. Stringer C. The origin and evolution of *Homo sapiens*. *Philos Trans R Soc Lond B Biol Sci.* 2016; 371.
51. Carroll RL, Kuntz A, Albright K. Vertebral development and amphibian evolution. *Evol Dev.* 1999; 1:36-48.
52. Venkatesh B, Lee AP, Ravi V, *et al.* Elephant shark genome provides unique insights into gnathostome evolution. *Nature.* 2014; 505:174-179.
53. Stemple DL. Structure and function of the notochord: An essential organ for chordate development. *Development.* 2005; 132:2503-2512.
54. Sibson R. SLINK: An optimally efficient algorithm for the single-link cluster method. *Comput J.* 1973; 16:30-34.
55. Jiao Y, Wickett NJ, Ayyampalayam S, *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature.* 2011; 473:97-100.
56. Innan H, Kondrashov F. The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet.* 2010; 11:97-108.
57. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. *J Genet.* 2013; 92:155-161.
58. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature.* 2007; 449:54-61.
59. Zhang L, Lu HH, Chung WY, Yang J, Li WH. Patterns of segmental duplication in the human genome. *Mol Biol Evol.* 2005; 22:135-141.
60. McCoy VE, Saupé EE, Lamsdell JC, *et al.* The 'Tully monster' is a vertebrate. *Nature.* 2016; 532:496-499.
61. Kugler JE, Kerner P, Bouquet JM, Jiang D, Di Gregorio A. Evolutionary changes in the notochord genetic toolkit: A comparative analysis of notochord genes in the ascidian *Ciona* and the larvacean *Oikopleura*. *BMC Evol Biol.* 2011; 11:21.
62. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform.* 2005; 6:118-134.
63. Wang L, Stein LD. Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process. *BMC Evol Biol.* 2013; 13:57.
64. Gelfman S, Burstein D, Penn O, Savchenko A, Amit M, Schwartz S, Pupko T, Ast G. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res.* 2012; 22:35-50.
65. Fuertes MA, Rodrigo JR, Alonso C. Do Intron and Coding Sequences of Some Human-Mouse Orthologs Evolve as a Single Unit? *J Mol Evol.* 2016; 82:247-250.
66. Clancy S. RNA splicing: Introns, exons and spliceosome. *Nature Education.* 2008; 1:31.
67. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: Diversification, exon definition and function. *Nat Rev Genet.* 2010; 11:345-355.
68. Gyekis J, Blizard DA, Stout JT, Vandenberg DJ, McClearn GE, Hager R. Genetic and maternal effects on offspring mortality in mice. *Evol Biol.* 2011; 38:434-440.
69. Simon AK, Hollander GA, McMichael A. Evolution of the immune system in humans from infancy to old age. *Proc Biol Sci.* 2015; 282:20143085.
70. Melville JM, Moss TJ. The immune consequences of preterm birth. *Front Neurosci.* 2013; 7:79.
71. Goto M. Inflammaging (inflammation + aging): A driving force for human aging based on an evolutionarily antagonistic pleiotropy theory? *Biosci Trends.* 2008; 2:218-230.

(Received December 31, 2016; Revised February 22, 2017; Accepted March 7, 2017)