[Cell Cycle 4:9, 1281-1285; September 2005]; ©2005 Landes Bioscience

Report

APOBEC4, a New Member of the AID/APOBEC Family of Polynucleotide (Deoxy)cytidine Deaminases Predicted by Computational Analysis

Igor B. Rogozin[1]
Malay K. Basu[1]
I. King Jordan[1]
Youri I. Pavlov[2-4]
Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information NLM; National Institutes of Health; Bethesda, Maryland USA;

[2]Eppley Institute for Research in Cancer; [3]Department of Biochemistry and Molecular Biology; [4]Department of Pathology and Microbiology; University of Nebraska Medical Center; Omaha, Nebraska USA

*Correspondence to: Eugene V. Koonin; National Center for Biotechnology Information NLM; National Institutes of Health; Bethesda, Maryland 20894 USA; Tel.: 301.435.5913; Fax: 301.435.7794; Email: koonin@ncbi.nlm.nih.gov

Received 07/01/05; Accepted 07/06/05

Previously published as a *Cell Cycle* E-publication:
http://www.landesbioscience.com/journals/cc/abstract.php?id=1994

KEY WORDS

cytidine deaminase, DNA/RNA modification, phylogenetic analysis, editing enzyme, innate immunity, APOBEC1

ABSTRACT

Using iterative database searches, we identified a new subfamily of the AID/APOBEC family of RNA/DNA editing cytidine deaminases. The new subfamily, which is represented by readily identifiable orthologs in mammals, chicken, and frog, but not fishes, was designated APOBEC4. The zinc-coordinating motifs involved in catalysis and the secondary structure of the APOBEC4 deaminase domain are evolutionarily conserved, suggesting that APOBEC4 proteins are active polynucleotide (deoxy)cytidine deaminases. In reconstructed maximum likelihood phylogenetic trees, APOBEC4 forms a distinct clade with a high statistical support. APOBEC4 and APOBEC1 are joined in a moderately supported cluster clearly separated from AID, APOBEC2 and APOBEC3 subfamilies. In mammals, APOBEC4 is expressed primarily in testis which suggests the possibility that it is an editing enzyme for mRNAs involved in spermatogenesis.

INTRODUCTION

Cytidine deaminases (CDAs; EC 3.5.4.5) catalyze the deamination of cytidine to uridine and are important in the pyrimidine salvage pathway in prokaryotes and eukaryotes. These enzymes contain a zinc-coordinating domain with the characteristic motif (H/C)xE … PCxxC (x stands for any residue). The Zn ion in the active site plays a central role in the proposed catalytic mechanism by activating a water molecule to form a hydroxide ion that performs a nucleophilic attack on the substrate.[1-3]

The cytidine deaminase superfamily also includes the AID/APOBEC family, which is a vertebrate-specific RNA/DNA editing expansion of deaminases. One of the best characterized modes of mRNA-editing is cytidine to uridine (C > U) deamination catalyzed by APOBEC1. APOBEC1 is the catalytic component of a complex that edits apolipoprotein B mRNA by catalyzing the C6666U deamination which creates a premature stop codon and causes tissue-specific production of a truncated apolipoprotein B polypeptide chain.[4,5] The AID/APOBEC protein family contains four subfamilies (AID, APOBEC1, APOBEC2, and APOBEC3) and includes several members with an experimentally confirmed capability of deaminating cytosine to uracil in single-stranded polynucleotides, while fulfilling diverse physiological functions.[6-8] AID functions in the adaptive humoral immune response, namely, somatic hypermutation of the immunoglobulin V gene and switch recombination of the immunoglobulin C gene.[9,10] Two members of the diverse APOBEC3 subfamily (human APOBEC3G and APOBEC3F) are involved in an innate pathway of restriction of retrovirus infection, presumably, by deaminating cytosines in viral first-strand cDNA replication intermediates[7,11,12] although more complex models have also been proposed.[13] The physiological functions of APOBEC2[14,15] and of other APOBEC3s are unknown.[7,8,12] However, the expansion and rapid evolution of APOBEC3 proteins in the primate lineage (there are eight human APOBEC3 genes) suggests that at least some of these proteins might have antiviral functions.[8]

APOBEC1 was the first member of the family to be discovered[4,5] and has become the paradigm for subsequent studies. However, phylogenetic analysis indicates that APOBEC1 is a recent evolutionary arrival whereas AID and APOBEC2 are the ancestral family members.[8] We analyzed distant similarities of AID/APOBEC protein sequences and identified a previously undetected subfamily of AID/APOBEC homologs which we provisionally named APOBEC4. Phylogenetic analysis suggests that this protein subfamily is most closely related to APOBEC1, however, it has a wider phyletic distribution similar to that of AID/APOBEC2 and, accordingly, appears to have emerged early in vertebrate evolution.

©2005 LANDES BIOSCIENCE. DO NOT DISTRIBUTE

MATERIALS AND METHODS

The non-redundant (nr) database of National Center for Biotechnology Information (http://ncbi.nlm.nih.gov/) and vertebrate genomes at the ENSEMBL web site (http://www.ensembl.org/) were searched using the BLASTP program.[16] Nucleotide genome sequences were searched using TBLASTN with protein sequences as queries. Iterative sequence similarity searches were performed using PSI-BLAST with a single sequence used as the query and with default parameters.[16] Each search was run for a minimum of three iterations or to convergence. Multiple alignments were generated using the MUSCLE program with 50 iterations.[17] The resulting multiple alignment was corrected manually using the PSI-BLAST results, the known three dimensional structures of CDAs, and predicted secondary structure of APOBECs as additional guides. Protein secondary structure was predicted using JPRED.[18]

Phylogenetic analyses were performed using minimum evolution (least-square) and maximum likelihood methods. To generate the input file, all columns containing gaps were either deleted pairwise or entirely deleted from the corrected alignment. Minimum evolution trees were constructed using either MEGA3[19] with the Poisson correction model and pairwise deletion of gaps and 1000 bootstrap replicates or using the FITCH program of the PHYLIP package[20] with 1000 bootstrap replicates, after complete removal of all gapped columns. Maximum likelihood trees were generated using a two-step procedure. At the first step, a minimum evolution tree was generated using FITCH, and at the second step, the topology of this tree was used as the input to PROTML[21] to produce a maximum likelihood tree using local rearrangements. The statistical significance of the internal nodes of the resulting maximum likelihood tree was then determined using relative estimate of logarithmic likelihood bootstrap (RELL-BP) as implemented in PROTML.[21] In the second method, an initial tree was constructed using the PROTML program of the PHYLIP package, with star decomposition. The tree topology was used as a guide to generate maximum likelihood trees using the PhyML program with 100 bootstrap replicates generated from the input alignment.[22] The consensus tree of these 100 bootstrapped trees was derived using the CONSENSE program of the PHYLIP package to obtain the bootstrapped the full maximum likelihood tree. Both methods produced phylogenetic trees with the same topology with respect to the main branching.

Table 1 **The APOBEC4 subfamily of predicted cytidine deaminases**

Species	Name	Refseq	GI	ESTs
Human	Hypothetical protein	NM_203454	44888831	8 testis, 1 brain, 1 uterus
Macaca fasicularis	Testicular cDNA	-	17026052	-
Mouse	Hypothetical protein	XM_355245	38073497	2 testis
Rat	Hypothetical protein	XP_573474	62945336	1 testis
Cow	Hypothetical protein	XP_613244	61875234	-
Chicken	Hypothetical protein	XP_426631	50751204	-
Xenopus tropicalis	Hypothetical protein GENSCAN00000137932	-	-	-

Table 2 **Number of amino acid substitutions per site in human and mouse orthologs of the AID/APOBEC family**

Protein	AID	APOBEC1	APOBEC2	APOBEC3	APOBEC4
Number of substitutions	0.09	0.31	0.10	1.03	0.30
Standard error	0.03	0.05	0.03	0.12	0.05

The number of amino acid substitutions and standard errors were calculated using the Poisson correction as implemented in MEGA3.[19] In the case of APOBEC3, the human APOBEC3F and mouse APOBEC3 proteins were used.

RESULTS AND DISCUSSION

Identification of APOBEC4 by sequence similarity searches. A PSI-BLAST search with the human AID sequence as the query returns hits to AID/APOBEC proteins in the first and second iterations. Unexpectedly, from the third iteration onward, this and other searches with AID/APOBEC sequences started to recover uncharacterized vertebrate proteins (Table 1). Reciprocal searches using these sequences as queries readily recovered the AID/APOBEC sequences, suggesting that these proteins and AID/APOBEC proteins are, indeed, homologs. We called this newly discovered protein subfamily APOBEC4. BLAST searches with the human APOBEC4 sequence (GI: 44888831) as a query readily identified the presence of the APOBEC4 gene in the *Xenopus tropicalis* genome. By contrast, these searches failed to identify APOBEC4 protein in any of the available (nearly) complete fish genomes.

Sequence and structural features of APOBEC4. To further examine the sequence and structural conservation between CDA, AID, APOBEC1, APOBEC2, APOBEC3, and APOBEC4, we constructed a multiple alignment of 4 CDA sequences and 45 AID/APOBEC sequences (Fig. 1). In the case of the dimeric APOBEC3 proteins,[8] the more conserved C-terminal domain, which determines the specificity of retrovirus hypermutation induced by human APOBEC3F and APOBEC3G,[23] was used for this analysis. The alignment shows notable conservation of the Zn-coordinating motif, (H/C)xE… $PCx_{2-6}C$ (Fig. 1). However, the proline residue, which is present in the middle of the HxE motif in most of the APOBEC4 sequences, aligns with an alanine in the CDA sequences whereas AID/APOBEC sequences contain several other amino acids in this position (Fig. 1). Interestingly, the *Xenopus tropicalis* APOBEC4 contains an alanine in the middle of the HxE motif (Fig. 1) suggesting that this could be the ancestral state of the motif. A distinctive feature of APOBEC4 is the insertion of four amino acids between the conserved cysteines of the PC…C motif; the presence of this insert complicates the detection of APOBEC4 in sequence similarity searches (see above). It should be noted that even longer inserts are present in this motif in several deaminases outside the CDA/AID/APOBEC superfamily,[24] which is compatible with the prediction that APOBEC4 is an active deaminase.

Structural homology models based on *E. coli* or yeast CDA structures have been previously proposed for APOBEC1, AID and APOBEC3G.[25-27] The deaminase domain of APOBEC4 conforms to the b1b2a1b3a2b4a3b5 arrangement (a designates an α-helix and b designates a β-strand) typical of the AID/APOBEC family rather than the b1b2a1b3a2b4b5 arrangement (with the a3 helix missing) seen in the CDAs (Fig. 1). The additional a3 helix is a signature of the AID/APOBEC family.[27] With the exception of this helix, the predicted secondary structural elements of APOBEC4 show a nearly perfect correspondence with the elements derived from the 3D structure of CDAs, supporting the notion that these proteins contain a domain with the same fold (Fig. 1).

Phylogenetic analysis and evolutionary implications. To gain insight into the evolution of the AID/APOBEC family and, in particular, the origin of APOBEC4, we constructed phylogenetic trees from the multiple alignment shown in Figure 1. The APOBEC1, APOBEC2 and APOBEC4 subfamilies each formed a clade with high statistical support (data not shown). Unexpectedly, the members of the APOBEC3 subfamily did not form a distinct clade but instead were interspersed with the AID proteins (data not shown) although previous phylogenetic analyses suggested monophyly of

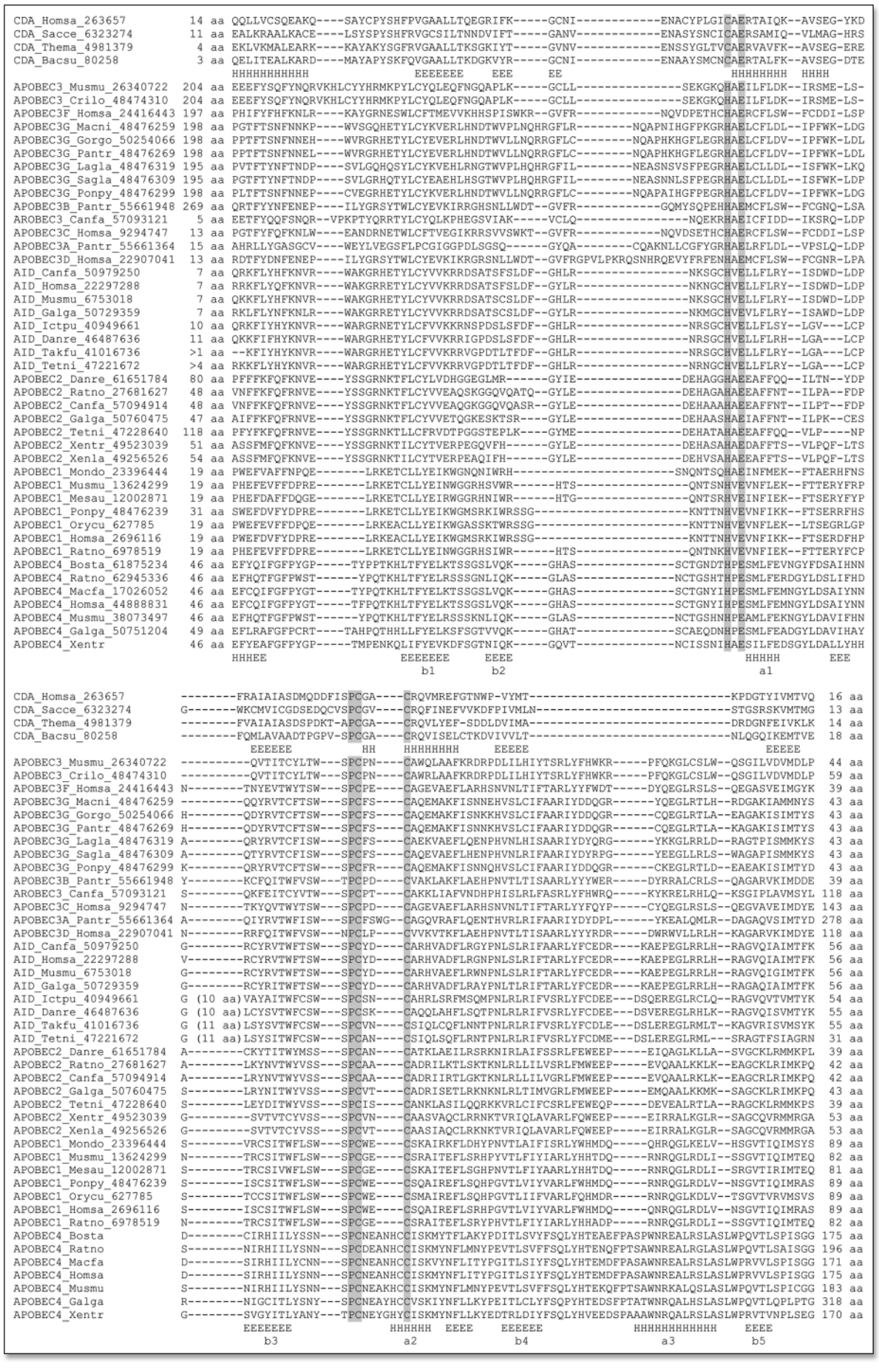

Figure 1 (previous page). Multiple alignment of the CDA/AID/APOBEC superfamily. Only the deaminase domain that is conserved in all proteins is shown. The proteins are designated by the corresponding species abbreviation appended by GI numbers. The secondary structure of the CDA proteins derived from the available crystal structures[27] is shown under the CDA sequences. The α-helices are denoted by "H"s and marked a1 and a2, and the β-strands are denoted by "E"s and marked b1–b5. The predicted secondary structure of the APOBEC4 subfamily is plotted below the alignment. Columns with 100% conserved residues are shaded gray. The species abbreviations are as follows: Bacsu, *Bacillus subtilis*; Bosta, *Bos taurus*; Canfa, *Canis familiaris*; Crilo, *Cricetulus longicaudatus*; Danre, *Danio rerio*; Galga, *Gallus gallus*; Gorgo, *Gorilla gorilla*; Homsa, *Homo sapiens*; Ictpu; *Ictalurus punctatus*; Lagla, *Lagothrix lagotricha*; Macfa, *Macaca fascicularis*; Macni, *Macaca nigra*; Mesau, *Mesocricetus auratus*; Mondo, *Monodelphis domestica*; Musmu, *Mus musculus*; Orycu, *Oryctolagus cuniculus*; Pantr, *Pan troglodytes*; Ponpy, *Pongo pygmaeus*; Ratno, *Rattus norvegicus*; Sacce, *Saccharomyces cerevisiae*; Sagla, *Saguinus labiatus*; Takru, *Takifugu rubripes*; Tetni, *Tetraodon nigroviridis*; Thema, *Thermotoga maritima*; Xenla, *Xenopus laevis*; Xentr, *Xenopus tropicalis*.

this subfamily.[8] The problems with the APOBEC3 subfamily are likely to be caused by long-branch attraction artifacts (long branches tend to corrupt the phylogenetic signal)[28] given the dramatic differences in the evolutionary rates between the fast-evolving APOBEC3 and other subfamilies (Table 2). We showed that APOBEC4 evolved at an intermediate rate, much lower than that of APOBEC3, similar to that of APOBEC1, but considerably greater than APOBEC2 and AID (Table 2).

To minimize long-branch attraction artifacts in tree reconstruction, we removed the fast-evolving APOBEC3 subfamily (see Table 2) from the present analysis. In the reconstructed tree, which can be rooted with the CDAs, APOBEC1 and APOBEC4 form distinct clades that are joined in a moderately supported cluster (Fig. 2). The APOBEC2 and AID subfamilies form a third clade with a high bootstrap support (Fig. 2). Interestingly, APOBEC4 is present in mammals, birds and amphibia (but so far not fishes) similarly to AID and APOBEC2 but unlike APOBEC1 which is so far restricted to mammals (Fig. 2 and Table 1). This suggests that the duplication leading to the distinct APOBEC1 and APOBEC4 genes might have occurred prior to the amphibia-reptile divergence, perhaps, with subsequent loss of APOBEC1 in some lineages. This interpretation contradicts the hypothesis that APOBEC1 is a mammalian-specific derivative of AID.[8] However, phylogenetic reconstructions for families with a high rate variation among subfamilies (Table 2) should be interpreted with utmost caution due to the substantial impact of long-branch attraction artifacts on tree topology.[28] In particular, it cannot be strictly ruled out that the APOBEC1-APOBEC4 clade is a result of such an artifact.

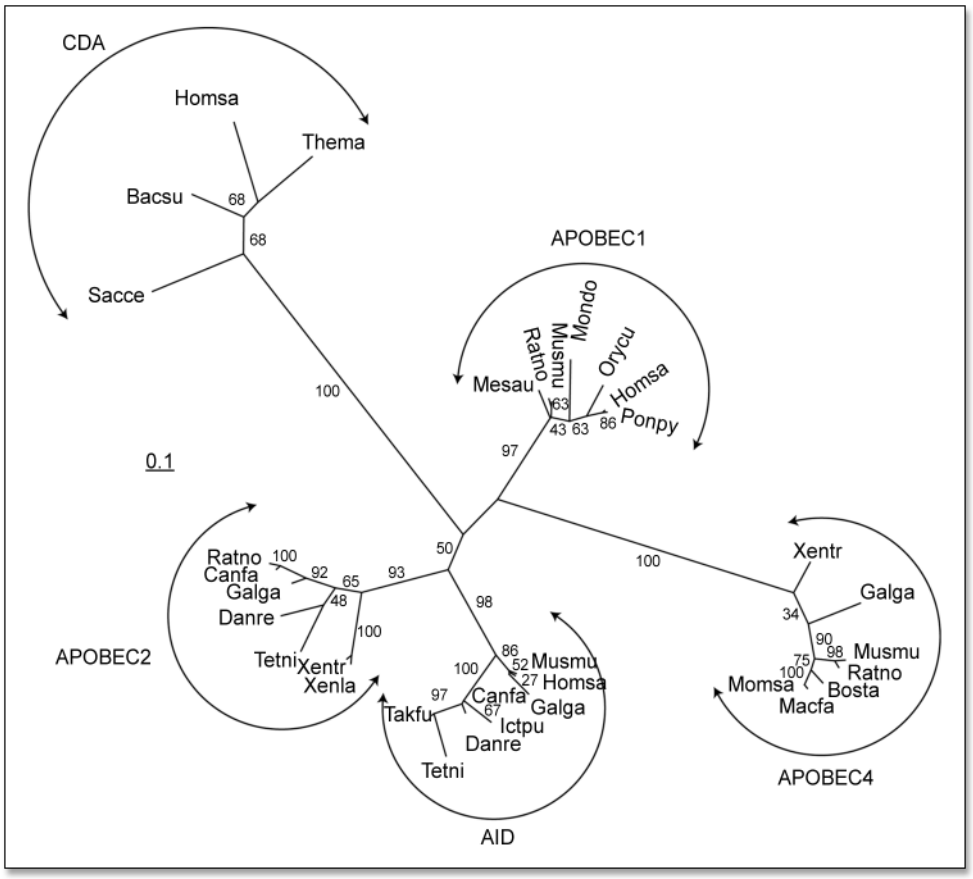

Figure 2. A maximum likelihood phylogenetic tree of the CDA/AID/APOBEC superfamily. Maximum likelihood RELL bootstrap support values are shown next to internal branches. The species abbreviations are as in Figure 1.

The deep internal branches of the tree, particularly, the branch connecting CDA and AID/APOBEC families, are very long compared to the branches within the CDA clade and within each clade of the AID/APOBEC family (Fig. 2). This suggests that the ancestor of AID/APOBEC family had an accelerated rate of evolution which still can be observed in the APOBEC3 subfamily (Table 2). Such acceleration of evolution characteristic of proteins involved in direct interactions with infectious agents (e.g., viruses)[29] which is compatible with the accelerated rate of evolution and the confirmed antiviral function of some APOBEC3 subfamily members.[7,11,12,30,31] Thus, suppression of infectious agents might be the original function of AID/APOBEC ancestors. Later in evolution, ancestors of AID, APOBEC1, APOBEC2 and APOBEC4 proteins gained functions different from the ancestral one and their evolution substantially slowed down (Table 2 and Fig. 2).

BIOLOGICAL IMPLICATIONS AND CONCLUSIONS

The spectrum of biological functions of the editing enzymes of the AID/APOBEC family is expanding. In particular, it has been recently shown that these deaminases, in addition to mRNA editing, are required for innate immunity to retroviruses and humoral immunity.[7,9-12] Here we describe a new subfamily of AID/APOBEC homologs, APOBEC4, which is represented by readily identifiable orthologs in mammals, chicken, and frog, but not fishes. The Zn-coordinating motifs and the secondary structure of the APOBEC4 deaminase domain are evolutionarily conserved which suggests that APOBEC4 proteins possess the polynucleotide (deoxy)cytidine deamination activity. Examination of mouse expression arrays (Fig. 3) and Expressed Sequence Tag (EST) data for human, mouse, and rat showed that APOBEC4 is expressed primarily in testis (Table 1). Based on this observation, it is tempting to speculate that APOBEC4 is an editing enzyme for mRNAs involved in spermatogenesis.

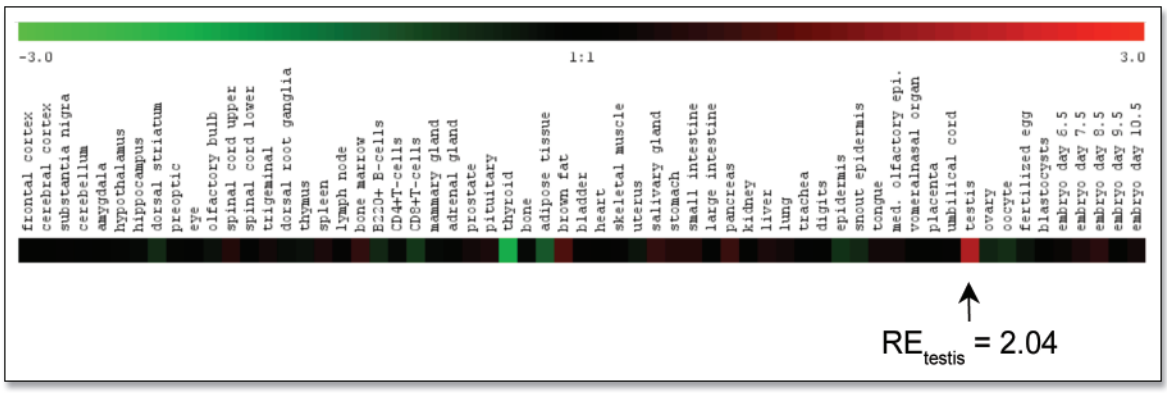

Figure 3. Preferential expression of mouse APOBEC4 in testis. The expression data are from the Novartis Gene Expression Atlas 2.[32] The figure shows a "thermal map" representation of the expression array data for APOBEC4 for the indicated mouse tissues. Colors are scaled to the relative expression level (REi) for each tissue (i). REi is equal to log2(Ei/M), where Ei is the signal intensity for each tissue and M is the median signal intensity for the gene. Tissue designations are given in the body of the figure. The >4x excess of APOBEC4 expression in testis over the median expression for all tissues was highly statistically significant according to one-sample t-test (p = 1.4e-36).

References

1. Cohen RM, Wolfenden R. Cytidine deaminase from *Escherichia coli*. Purification, properties and inhibition by the potential transition state analog 3,4,5,6-tetrahydrouridine. J Biol Chem 1971; 246:7561-5.
2. Munch-Petersen A, Nygaard P, Hammer-Jespersen K, Fiil N. Mutants constitutive for nucleoside-catabolizing enzymes in *Escherichia coli* K12. Isolation, charactrization and mapping. Eur J Biochem 1972; 27:208-15.
3. Carter Jr CW. The nucleoside deaminases for cytidine and adenosine: Structure, transition state stabilization, mechanism, and evolution. Biochimie 1995; 77:92-8.
4. Navaratnam N, Morrison JR, Bhattacharya S, Patel D, Funahashi T, Giannoni F, Teng BB, Davidson NO, Scott J. The p27 catalytic subunit of the apolipoprotein B mRNA editing enzyme is a cytidine deaminase. J Biol Chem 1993; 268:20709-12.
5. Teng B, Burant CF, Davidson NO. Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science 1993; 260:1816-9.
6. Durandy A. Activation-induced cytidine deaminase: A dual role in class-switch recombination and somatic hypermutation. Eur J Immunol 2003; 33:2069-73.
7. Pham P, Bransteitter R, Goodman MF. Reward versus risk: DNA cytidine deaminases triggering immunity and disease. Biochemistry 2005; 44:2703-15.
8. Conticello SG, Thomas CJ, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol 2005; 22:367-77.
9. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell 2000; 102:553-63.
10. Revy P, Muto T, Levy Y, Geissmann F, Plebani A, Sanal O, Catalan N, Forveille M, Dufourcq-Labelouse R, Gennery A, Tezcan I, Ersoy F, Kayserili H, Ugazio AG, Brousse N, Muramatsu M, Notarangelo LD, Kinoshita K, Honjo T, Fischer A, Durandy A. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell 2000; 102:565-75.
11. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 2002; 418:646-50.
12. Neuberger MS, Harris RS, Di Noia J, Petersen-Mahrt SK. Immunity through DNA deamination. Trends Biochem Sci 2003; 28:305-12.
13. Sasada A, Takaori-Kondo A, Shirakawa K, Kobayashi M, Abudu A, Hishizawa M, Imada K, Tanaka Y, Uchiyama T. APOBEC3G targets human T-cell leukemia virus type 1. Retrovirology 2005; 2:32.
14. Liao W, Hong SH, Chan BH, Rudolph FB, Clark SC, Chan L. APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. Biochem Biophys Res Commun 1999; 260:398-404.
15. Anant S, Mukhopadhyay D, Sankaranand V, Kennedy S, Henderson JO, Davidson NO. ARCD-1, an apobec-1-related cytidine deaminase, exerts a dominant negative effect on C to U RNA editing. Am J Physiol Cell Physiol 2001; 281:C1904-16.
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-402.
17. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792-7.
18. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999; 292:195-202.
19. Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 2004; 5:150-63.
20. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996; 266:418-27.
21. Hasegawa M, Kishino H, Saitou N. On the maximum likelihood method in molecular phylogenetics. J Mol Evol 1991; 32:443-5.
22. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003; 52:696-704.
23. Hache G, Liddament MT, Harris RS. The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain. J Biol Chem 2005; 280:10920-4.
24. Dance GS, Beemiller P, Yang Y, Mater DV, Mian IS, Smith HC. Identification of the yeast cytidine deaminase CDD1 as an orphan C >U RNA editase. Nucleic Acids Res 2001; 29:1772-80.
25. Scott J, Navaratnam N, Carter C. Molecular modelling and the biosynthesis of apolipoprotein B containing lipoproteins. Atherosclerosis 1998; 141(Suppl 1):S17-24.
26. Navaratnam N, Fujino T, Bayliss J, Jarmuz A, How A, Richardson N, Somasekaram A, Bhattacharya S, Carter C, Scott J. *Escherichia coli* cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. J Mol Biol 1998; 275:695-714.
27. Huthoff H, Malim MH. Cytidine deamination and resistance to retroviral infection: Towards a structural understanding of the APOBEC proteins. Virology 2005; 334:147-53.
28. Philippe H. Opinion: Long branch attraction and protist phylogeny. Protist 2000; 151:307-16.
29. Hughes AL. Adaptive evolution of genes and genomes. Oxford: Oxford Univ, 1999.
30. Zhang J, Webb DM. Rapid evolution of primate antiviral enzyme APOBEC3G. Hum Mol Genet 2004; 13:1785-91.
31. Sawyer SL, Emerman M, Malik HS. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol 2004; 2:E275.
32. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 2004; 101:6062-7.