Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity

Arun Kumar Krishnan*, Lakshminarayan M. Iyer*, Stephen J. Hollanda, Thomas Boehmb, and L. Aravinda,b,1

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; and bDepartment of Developmental Immunology, Max Planck Institute of Immunobiology and Epigenetics, 79108 Freiburg, Germany

Edited by Anjana Rao, La Jolla Institute and University of California San Diego, La Jolla, CA, and approved February 23, 2018 (received for review November 30, 2017)

AID/APOBEC deaminases (AADs) convert cytidine to uridine in single-stranded nucleic acids. They are involved in numerous mutagenic processes, including those underpinning vertebrate innate and adaptive immunity. Using a multipronged sequence analysis strategy, we uncover several AADs across metazoa, dictostelid, and algae, including multiple previously unreported vertebrate clades, and versions from urochordates, nematodes, echinoderms, arthropods, lophotrochozoans, cnidarians, and porifera. Evolutionary analysis suggests a fundamental division of AADs early in metazoan evolution into secreted deaminases (SNADs) and classical AADs, followed by diversification into several clades driven by rapid-sequence evolution, gene loss, lineage-specific expansions, and lateral transfer to various algae. Most vertebrate AADs, including AID and AP0BEC1–3, diversified in the vertebrates, whereas the APOBEC4-like clade has a deeper origin in metazoa. Positional entropy analysis suggests that several AAD clades are diversifying rapidly, especially in the positions predicted to interact with the nucleic acid target motif, and with potential viral inhibitors. Further, several AADs have evolved neomorphic metal-binding inserts, especially within loops predicted to interact with the target nucleic acid. We also observe polymorphisms, driven by alternative splicing, gene loss, and possibly intergenic recombination between paralogs. We propose that biological conflicts of AADs with viruses and genomic retroelements are drivers of rapid AAD evolution, suggesting a widespread presence of mutagenesis-based immune-defense systems. Deaminases like AID represent versions “institutionalized” from the broader array of AADs pitted in such arms races for mutagenesis of self-DNA, and similar recruitment might have independently occurred elsewhere in metazoa.

The deaminase superfamily encompasses zinc-dependent enzymes catalyzing the deamination of bases in free nucleotides and nucleic acids across diverse biological contexts (1–3). These include enzymes that are (i) primarily involved in the salvage strategy of bases in free nucleotides, such as the cytidine deaminases (CDD/CDA) (4), deoxyxytidylate monophosphate deaminases (dCMP) (5), and guanine deaminase (GuaD) (6); (ii) engaged in the biosynthesis of modified nucleotide-derived compounds (e.g., blastcidin-S deaminase) (7) and the RibD deaminase domain (8); and (iii) active in situ modifications of bases in nucleic acids, such as tRNA adenine deaminases (Tad2/TadA and Tad1) (9), RNA-specific adenosine deaminase (ADAR) (10), RNA-editing cytide deaminase (ADAR) (11), DNA-editing cytide deaminases such as DYW (11, 12), and members of the AID (activation-induced deaminase)/APOBEC family, which modify both DNA and RNA (13, 14). This last group of deaminases is implicated in the generation of degenerate codons for decoding during translation (Tad2/TadA) (9, 15); stabilization of codon-anti-codon interactions (Tad1) (16, 17); editing mRNAs, siRNAs, and miRNA precursors; inactivation of RNA viruses by hypermutation (ADAR) (10, 18, 19); diversification of antibodies (AID) (20); and defense against retroviruses and retrotransposons through hypermutation of RNA during reverse transcription (APOBEC3s) (21).

The deaminase superfamily displays a conserved β-sheet with five β-strands arranged in 2-1-3-4-5 order interleaved with three α-helices forming an α/β-fold (the deaminase fold) (22), which it shares with JAB/RadC, AICAR transformylase, formate dehydrogenase accessory subunit (FdhD), and Tm1506 superfamilies of proteins. The active site consists of two zinc (Zn)-chelating motifs, respectively typified by the signatures HxE/CxE/DxE at the end of helix 2 and Cx,C (where x is any amino acid and n ≥ 2) located in loop 5 and the beginning of helix 3 (Fig. 1). The deaminase superfamily contains two major divisions. In the so-called helix-4 division, which includes the Tad2/TadA, ADAR, and AID/APOBEC-like deaminases (AADs), the helix 4 precedes the terminal strand, resulting in strand 5 being parallel to the rest of the sheet (Fig. 1). In the C-terminal hairpin division, containing families such as the CDD-like, blastcidin-S-like, and DYW, strands 4 and 5 immediately follow each other, forming a β-hairpin (22).

The best-known AADs are APOBEC1, AID, APOBEC2, APOBEC3s (3A–3D; 3F–3H), and APOBEC4 (13). AID plays a critical role in gnathostome (jawed vertebrate) adaptive immunity.

Significance

Mutagenic AID/APOBEC deaminases (AADs) are central to processes such as generation of antibody diversity and antiviral defense in vertebrates. Their presence and role outside vertebrates are poorly characterized. We report the discovery of several AADs, including some that are secreted, across diverse metazoan, dictostelid, and algal lineages. They appear to have emerged from an early transfer of an AAD from bacterial toxins systems, followed by extensive diversification into multiple eukaryotic clades, showing dramatic structural innovation, rapid divergence, gene loss, polymorphism, and lineage-specific expansions. We uncover evidence for their divergence in arms-race scenarios with viruses and genomic retroelements and show that AAD-based nucleic acid mutagenesis as a basis of immune defense is widespread across metazoa, slime molds, and algae.

Author contributions: A.K., L.M.I., S.J.H., T.B., and L.A. designed research; A.K., L.M.I., and L.A. performed research; S.J.H., T.B., and L.A. contributed new reagents/analytic tools; A.K., L.M.I., and L.A. analyzed data; and A.K., L.M.I., and L.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The sequence reported in this paper has been deposited in the Sequence Read Archive at the NCBI (accession nos. SAMP08013505, SAMP08013506, and SAMP08013507).

To whom correspondence should be addressed. Email: aravind@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1720897115/-/DCSupplemental.

Published online March 19, 2018.

www.pnas.org/cgi/doi/10.1073/pnas.1720897115
Fig. 1. Characterization of the AAD protein family. (A) MSAs of representatives of AAD clades labeled with accessions, clade names, and species abbreviations (full species names in SI Appendix, Fig. S1). Zn-catalyzing active site residues are highlighted in black, and known substrate-interacting residues are shown in red. The SNAD1/2-specific helix between strands 1 and 2 and neomorphic strands are shown in gray rectangles. (B) Topology diagram depicting the conserved core common to all AADS, the nucleic acid substrate, and key residues. (C) Topologies of clades showing Zn-chelating features. Homologous Zn-chelating residues between APOBEC4 and the Cnidaria-algae clade are highlighted in purple, green, and blue, whereas other Zn-chelating residues are colored gray. (D) Topologies of SNAD1–3: Inserts embedded within the core domain are in blue, while N- and C-terminal extensions are in gray.

cytidine deamination by AID at the genomic Ig loci in mature B cells triggers base excision and mismatch repair mechanisms (20, 23–27). This process is central to class-switch recombination (24). Additionally, at least in some jawed vertebrates, AID-catalyzed deamination causes somatic hypermutation and/or gene conversion, which drives antibody diversification, for instance as part of affinity maturation during an immune response (28, 29). The AADS expressed in a gamathec (jawless vertebrate) lymphocytes (PmCD41 and PmCD42) are
implicated in gene conversion-based diversification of their antigen receptors (20), the variable lymphocyte receptors (VLRs), which are structurally unrelated to antibodies (review ref. 30). APOBEC3 deaminases act as restriction factors in the innate response to retroviruses, herpesviruses, and parasites. They are also involved in the defense against retroviruses and genetic retrotransposons (36, 37). APOBEC2 and APOBEC4 remain poorly understood in terms of their molecular functions and substrate specificity.

Interestingly, several deaminase clades catalyzing organellar RNA-editing and DNA mutation, including the AADs, are sporadically distributed across the tree of life. Our previous analysis revealed that one of the main factors for this pattern is the diversification of most deaminase families as toxic effectors in bacterial toxic systems, followed by multiple independent lateral transfers to eukaryotes (22). Whereas the evolutionary history of AADs in vertebrates has been extensively examined (22, 38), our prior discovery of AADs (22) outside vertebrates, and sporadically in other eukaryotic lineages and bacteria, hinted at the possible presence of as-yet-undiagnosed divergent homologs. In this study, a multipronged sequence analysis approach uncovered multiple clades of AADs from across metazoa, dictyosteliida, and algae. This allowed us to develop a comprehensive evolutionary picture of AAD diversification. We hypothesize that biological conflicts with viruses and genomic retroelements are the primary selective force behind the evolution of multiple independent lineage-specific expansions (LSEs) of these deaminases. Furthermore, in some cases, the genomic organization might favor polymorphism via recombination between paralogs, duplication, and gene loss. Our study suggests that the self-DNA–mutating deaminases, like AID and the cyclodeamine deaminases, emerged via “institutionalization” of AADs that were originally deployed in biological conflict with selfish genomic elements.

Results and Discussion

Identification and Classification of Members. In an earlier analysis, we discovered AADs in several bacteria and eukaryotic taxa other than vertebrates (22). This sporadic distribution, coupled with rapid sequence divergence, hinted at the possible presence of additional AAD versions that had evaded detection. Taking advantage of the several new genome sequences that were published since our last study, we developed a strategy of transitive and iterative sequence-profile searches against the National Center for Biotechnology Information (NCBI) nonredundant (nr) and Ensembl proteome databases to discover novel deaminase sequences. We ran PSI-BLAST and JALOUCHER and translating searches with TBLASTN. To isolate AADs from other deaminases, the newly detected sequences were analyzed by Profile-profile comparisons using the HHpred program, by examination of bidirectional best hits and by use of phylogenetic methods, and finally assessed for sequence and structural synapomorphies. Thus, we established a comprehensive collection of AADs from vertebrates, invertebrates, echinoderms, lophophorans, cnidarians, and algae, and added homologs from species within taxa that were previously known to possess AADs (Fig. L4 and SI Appendix, Figs. S3–S20), which we then used to identify shared sequence features and to compute phylogenetic trees and phyletic relationships. The overall phylogenetic tree strongly established the monophyly of a core group of gnathostome AADs, namely AID, APOBEC3, APOBEC2, APOBEC1, and the related lamprey CDAs (discussed in detail in the companion paper). Additionally, we identified two distinct clades of AADs that we designate NAD1 and NAD2 (Novel AID/APOBEC-like Deaminases 1 and 2; Figs. L4 and 24). These, too, group with the above AADs, specifically with the other gnathostome AADs. We refer to this entire assemblage of deaminases as the core vertebrate AAD clade. Our further analysis clarified the evolutionary trajectories and relationships for both the previously known AADs and the newly detected members (Fig. 24). First, APOBEC1 was retrieved in taxa such as birds, reptiles, amphibians, and lungfish, suggesting that its origin predated the tetrapod-lungfish divergence (SI Appendix, Figs. S2 and S4). Second, APOBEC2 was retrieved from choanoflagellans (sharks), actinopterygiens (ray-finned fishes), and sarcopterygiens (lungfish and coelacanth), thereby indicating its emergence before the divergence of gnathostomes (SI Appendix, Figs. S2 and S5). Third, multiple paralogous copies or splice forms of APOBEC1 were observed in some turtles and of AID in certain actinopterygiens and amphibians (SI Appendix, Figs. S2–S4). Fourth, APOBEC4 appeared to have emerged from AID in eutherian mammals, followed by paralog expansions.

The remaining AADs, from invertebrates such as urochordates (albeit predicted to be catalytically inactive), echinoderms, arthropods, lophophorans, and cnidarians, and APOBEC4 form multiple distinct clades that are out-groups to the above core vertebrate AAD clade (Fig. 24). We refer to the clade that unites all these AADs with core vertebrate AADs as the classical AAD clade (Fig. 24). Of these, APOBEC4 is recovered from tetrads, sarcopterygiens, and agnathans, but is frequently lost across actinopterygiens (Fig. 24 and SI Appendix, Figs. S2 and S3). Based on observed sequence features and a long, conserved metal-binding insert (Fig. 1C, Right panels, and SI Appendix, Fig. S1), we unified APOBEC4 with a specific group of AADs found in cnidarians and sporadically in phylogenetically distant algal lineages. This cnidaria-algae-APOBEC4 clade appears to be the outermost clade of the classical AADs. Its phyletic pattern suggests an origin early in metazoan evolution followed by transfer to photosynthetic eukaryotes possibly via the route of the algal-cnidarian symbioses. Another, remarkable, large monophyletic clade of AADs from diverse vertebrates and sponges contains members predicted to be catalytically inactive (SI Appendix, Figs. S2 and S13). Based on observed sequence features, we propose that one of these may be a long, conserved metal-binding insert (Fig. 1C, Right panels, and SI Appendix, Fig. S14). This clade formed a sister group to the classical AADs. Overall, this pattern suggests that there was a split between the classical AADs and the SNADs at the base of metazoan evolution (Fig. 24). We also detected secreted AADs in the dictyostelid slime molds Dictyostelium discoideal, Tegthnosporeula lacteum, and Ancylostoma, which might either group with the SNADs or form a distinct basal clade of eukaryote-specific AADs (Fig. 24 and SI Appendix, Fig. S4) and describe below the clades in greater detail (see SI Appendix for obtaining the complete set of sequences).

Members of Classical AAD Clade. Of the AADs, NAD1 and NAD2 clades belong to the core vertebrate AID/APOBEC clade but are not particularly close to any one of the known lineages (Fig. 24). Of these, NAD1 is found in ray-finned fishes, the coelacanth, amphibians, lizards, and marsupials, whereas NAD2 is restricted to amphibians (SI Appendix, Figs. S2, S7, and S8). NAD1 is typically found in a single copy per genome and has been independently lost in eutherian mammals and archosaurs. Among the classical invertebrate AADs, despite the patchy phyletic pattern suggestive of extensive gene loss, we delineated multiple well-defined clades, namely the following (Figs. 1 and 24): (i) a clade found in cnidarians, lophophorans, echinoderms, and tunicates with a deaminase domain most closely related to the core vertebrate AADs

Krishnan et al.
(SI Appendix, Fig. S9); (ii) a clade found in lophophorozoans and echinoderms with a metal-celathing insert involving elements from the extended loops 1 and 7 (SI Appendix, Fig. S10); and (iii) lineage-specific clades containing the versions from nematodes and arthropods, respectively (Fig. 2A and SI Appendix, Figs. S11 and S12).

The invertebrate classical AADs are dominated by LSEs, ranging from 2 to 22 paralogs per taxon (Fig. 2B and SI Appendix, Figs. S2 and S21). These numbers can widely differ within a single phylum; for example, among molluscs, the Japanese oyster *Crassostrea gigas* shows a large expansion of 22 distinct AADs, whereas *Biomytilus glabrata* appears to possess only a single representative. In general, several species of lophophorozoa and echinodermata display LSEs of classical AADs, whereas those arthropods that possess AADs typically encode a single or a few copies. Among cnidarians with AADs, approximately one-half contain only a single copy, while the other show LSEs. The prevalence of LSEs among invertebrate lineages parallels the APOBEC3 paralog expansion in mammals.

The Secreted AADs. The secreted AADs from metazoa can be divided into four major subclades, SNAD1–4, SNAD1, SNAD2, and SNAD3 clades are specific to vertebrates, whereas SNAD4 is found only in sponges (Fig. 2A and B and SI Appendix, Figs. S2 and S15–S18). In vertebrates, SNAD1 shows the widest phyletic spread, being present in all vertebrate groups. Among chondrichthyes, the platypus, the coelacanth, and actinopterygians, whereas SNAD2 and SNAD3 are only found in actinopterygians. In tetrapods, the two ancient clades SNAD1 and SNAD2 contain only a single copy, whereas SNAD3 and SNAD4 are present in a variety of lineages.

Hence, in vertebrates, SNAD1 is likely the ancestral clade that goes back to at least the stem euteleostomi. While SNAD2 and SNAD3 were derived from it as actinopterygian-specific expansions, SNAD4 showed a more limited distribution, confined to actinopterygians and mammals. SNAD4 clades are specific to terrestrial mammals, whereas SNAD2 and SNAD3 are found in all vertebrates. The prevalence of LSEs among invertebrate lineages parallels the APOBEC3 paralog expansion in mammals.

These unique structural features of the SNADs, combined with LSEs and differential loss of paralogous genes, indicate rapid and extensive divergence from the ancestral AAD archetype. The SNADs along with the AADs we detected in dictyostelids are unique among the eukaryotic deaminase lineages that are secrete. The secreted versions resemble the ancestral AADs from bacterial polyomorphytic toxins (22, 39), which were the building blocks of eukaryotic AADs. Another functional analogy is offered by the nucleic-acid-targeting CR (Crikler-RHS-type) effectors, which are secreted by various eukaryotes and delivered into recipient cells (40). The SNADs and dictyostelid AADs might be delivered into virus-infected cells or cells of extracellular parasites/pathogens. In vertebrates, the SNADs show a striking pattern of gene loss in lineages that reliably maintain a constant high body temperature, namely birds, marsupials, and placental mammals. By contrast, they are present in the basal members of these lineages that are either poikilothermic or have lower body temperatures (41).

**Cytosine in Single-Stranded Nucleic Acids Is the Likely Substrate Across the AAD Clade.** We used functional data for the core vertebrate AADs and the more distantly related tRNA-modifying TadA as models to understand the structure-function relationships that are conserved across the AADs. Across the deaminase superfAMILY, nucleic acid and nucleotide substrates are similarly positioned with the target base inserted into a binding pocket lined by helices 2 and 3 in proximity to the catalytic Zn2+ ion (Figs. 1B and 3A). In both TadA [Protein Data Bank (PDB): 2b3j] and APOBEC3A (PDB: 5oww), the target base is flipped out into the active site, away from the rest of the single-stranded nucleic acid backbone that is positioned in a U-shaped conformation (42, 43) (Fig. 3A). In both cases, the base 5′ to the target base (i.e., −1 position) is also flipped out, whereas the 3′ base (+1 position) of TadA also shows an altered conformation. These flanking bases might make contacts with the binding pocket and residues from loop 1 (between helix 1 and strand 1), loop 3 (between strand 2 and helix 2), loop 5 (between strand 3 and helix 3; in APOBEC3A), the end of strand 4, loop 7 (between strand 4 and helix 4), and the terminal helix (helix 5; in TadA).

Side chains of the following residues make specific contacts with the target base (Fig. 3A): (i) the active site histidine (PDB:
Deamination reactions; (ii) an asparagine (N57 in APOBEC3A; N42 in TadA) at the beginning of loop 3 makes polar contacts and might further interact with the phosphate backbone of the target nucleotide (as in APOBEC3A); and (iii) a tyrosine (Y130, in APOBEC3A) in loop 7 projects into the active site pocket and hydrogen-bonds with the backbone phosphate and also makes a π–π stacking interaction with the target nucleotide cytosine, as previously predicted (43). Analysis of the variability of these residues suggests that, while the histidine is absolutely conserved in active AADs, the asparagine in loop 3 might be replaced by other small residues such as serine, aspartate, or proline, and in some cases, a histidine residue, which can still play equivalent roles in contacting the cytosine. The tyrosine in loop 7 might be replaced by phenylalanine or tryptophan, hence retaining the aromatic character at this position. This configuration is a unique and strongly conserved feature across the AADs (Figs. 4 and 3A and E). The aromatic residue in loop 7 is missing in TadA, with its place taken by a smaller residue (e.g., D104 in TadA; PDB: 2b3j). As we previously predicted (22), an aromatic residue limits the size of the active site, such that only a pyrimidine base can fit into the pocket. By contrast, small residues at this position and other substrate-contacting positions in loop 7 of TadA result in a larger substrate-binding pocket accommodating an adenine (42). Additional residues in loop 3 and 7 of the targeted base make essential polar interactions with the polypeptide backbone, or interact with the sugar-phosphate backbone of the nucleic acid (Fig. 3A). One of these is the position at the end of strand 4, which is a strongly conserved basic residue (R128; PDB: 5sw) in the core vertebrate AADs, among others. It makes a polar contact with the backbone phosphate and possibly a cation–π interaction with the target base. However, substitutions at these positions are unlikely to drastically alter the size of the substrate pocket and the specificity for the target base. Based on these considerations, we propose that most AADs target a cytosine for deamination.

Inspection of the TadA structure reveals two residues: E45 (PDB: 2b3j) in loop 3 and D104 (PDB: 2b3j) in loop 7 that hydrogen-bond to the 2′ hydroxyl group of the ribose sugar. Indeed, a recent mutagenesis study has shown that alteration of D104 (PDB: 2b3j) to asparagine increases efficiency of adenine deamination in DNA substrates with respect to WT TadA (44). Hence, loops 3 and 7 might be involved in RNA vs. DNA discrimination. However, residues at these two positions, although often enriched in polar residues, are not universally conserved even across the TadAs (22), let alone the AADs. Further, in AADs, the equivalent of D104 is instead involved in pyrimidine selectivity as opposed to nucleic acid backbone interaction. This indicates that there are unlikely to be any widely conserved specificity determinants for the type of nucleic acid, offering an explanation as to why APOBEC1 (36) or TadA (45) targets both DNA and RNA substrates. In addition to clade-specific determinants, extraneous factors, such as the tissue of expression, interaction with other proteins, and subcellular localization, might play a key role in determining the type of nucleic acid that is modified.

Evidence for Differential Action of Selective Forces Across the AAD Clade. We used global and local position-specific Shannon entropy analysis to measure sequence variability and infer the possibility of purifying or positive selection on different AAD clades. We combined this information with published structures of AADs in complex with their substrates to interpret the observed variations (Fig. 3 A–D and SI Appendix, Fig. S22). Global entropy values showed that AID, APOBEC2, and APOBEC4 are slowly evolving, with AID being the slowest evolving of the three. In contrast, APOBEC3s, while originally derived from the more ancient AID (22), are evolving fast. High global sequence entropy can be associated with a role in an arms race with parasites (Fig. 3 C and D and SI Appendix, Fig. S22). APOBEC3 paralogs and APOBEC1, which are known to be deployed against viruses and retroelements, show high sequence variability, in turn an indicator of selective pressure arising from constantly evolving parasite resistance against them. In contrast, although AID is also involved in biological conflicts (i.e., immunity), it does not directly act on viral nucleic acids. Instead, it mutates the DNA of the host itself as part of the conserved antibody-diversification process in gnathostomes. Thus, AID seems to be under purifying rather than positive selection compared with the APOBEC3 and APOBEC1 proteins.

These observations serve as comparative yardsticks for the contrasting selective forces that have acted on different clades descending from the ancestral member of the AAD superfamily. Although there has been no direct experimental demonstration of the biochemical roles of APOBEC4 and APOBEC2, we find that, like AID, they show low global sequence entropy. Mutational data support a role for the deaminase active site of APOBEC2 in zebrafish retina (46) and mouse muscle (47) development. Hence, these AADs are likely involved in a conserved nucleic acid-editing function, possibly during development. With respect to the identified clades, we found that three groups of AADs show high global sequence entropy comparable to or even higher than APOBEC1 and APOBEC3:

i) The SNADs: Here, high variability is present across the sequence with the greatest diversity within the loops implicated in substrate binding and specificity (Fig. 3 C–E and SI Appendix, Fig. S22). Of note is the widespread substitution of the conserved CCA motif, whereas APOBEC3A prefers thymine at the −2 position (43, 48). Keeping with their lower global entropy values, the SNADs show high sequence diversity at these positions (43, 48). Keeping with their lower average global entropy, the entropy of substrate-binding positions is close to zero in AID, APOBEC2, and APOBEC4 (Fig. 3E). In contrast, APOBEC1, APOBEC3, most invertebrate classical AAD clades, and the SNADs show high sequence entropy at these positions (Fig. 3D and E and SI Appendix, Fig. S22). Of note is the frequently observed change in residue character at these positions within the same clade. These observations serve as comparative yardsticks for the contrasting selective forces that have acted on different clades descending from the ancestral member of the AAD superfamily. Although there has been no direct experimental demonstration of the biochemical roles of APOBEC4 and APOBEC2, we find that, like AID, they show low global sequence entropy. Mutational data support a role for the deaminase active site of APOBEC2 in zebrafish retina (46) and mouse muscle (47) development. Hence, these AADs are likely involved in a conserved nucleic acid-editing function, possibly during development. With respect to the identified clades, we found that three groups of AADs show high global sequence entropy comparable to or even higher than APOBEC1 and APOBEC3:

ii) The NAD1 clade found in several vertebrates.

iii) Most invertebrate classical AAD clades; these include clades exhibiting LSEs (average entropy values range between 2.0 and 2.5) and those from low-copy-number lineages (e.g., arthropods; average entropy = 2.4). This pattern indicates a function(s) in antiviral response, as is known for APOBEC1 and APOBEC3, with their divergence resulting from a comparable arms-race scenario in these biological conflicts.
Fig. 3. (A) Illustration of human APOBEC3A bound to ssDNA substrate (PDB ID code 5ssw) showing key APOBEC3A residues involved in Zn chelation and interactions with the target and neighboring bases (labeled −2, −1, +1, and +2). (B) Surface view of APOBEC3F (PDB: 4j4j) depicting Vif1-binding residues obtained from various studies. Buried residues: red; solvent exposed: blue. (C) Positional entropy values for various AAD clades with mean entropy values (blue horizontal lines) and secondary structures on top. Loop-1 and -7 residues are shown in E: blue dots. Refer to SI Appendix, Fig. S22, for details. (D) Boxplots comparing global entropy values. Mean entropy values of >2, ≤2, and ≤1.5 are colored deep-brick, blue, and green, respectively. (E) Sequence logos of key substrate-binding residues in loops 7 and 1. Loop-7 residues are equivalents of APOBEC3A (PDB: 5ssw) Y130, D131, and Y132. Loop-1 residues are equivalents of H29, K30, and T31.

as follows: (i) a previously undelineated metal-chelating insert in APOBEC4 supported by two cysteines and a histidine from loop 1 and a cysteine from loop 3; (ii) a metal-binding insert within loop 1 of several cnidian and algal sequences, which is a conserved feature of this clade; (iii) an insert stabilized by a metal-chelating histidine from loop 1 and three further Cys/His residues from loop 7 in the lophotrochozoan-echinoderm clade; and (iv) multiple disulfide bond-stabilized loops or extensions in vertebrate SNAD
and dictyostelid AAD clades. Together, these structures join the previously reported weak Zn-binding dyad of histidines in some members of the APOBEC3 clade (49). Despite differences in the chelating residues and locations of these clade-specific inserts, examination of their inferred positions suggests that they are proximal to the protein-nucleic acid interface.

Comparisons across AADs show that of the four substrate-binding loops, loop 3 shows the greatest variations in lengths across different clades, followed in order by loops 1, 7, and 5 (Fig. 4). Clades with low global sequence entropy (e.g., AID, APOBEC2, and APOBEC4) also show little variability in loop lengths. It has been previously reported that the length of loop 3 of APOBEC3 varies considerably (50); our dataset catalogs a range between 6 and 20 residues, with loop lengths of 6, 9, or 13 being most represented. Several other clades with high global entropy, such as the SNADs, the dictyostelid secreted AADs, and multiple invertebrate classical AADs, also show much loop-length variation. For example, the SNAD2 loop 3 varies between 3 and 64 residues, with ≥2 sequences possessing a loop longer than 10 residues and 29 sequences with lengths >15 residues (Fig. 4). Similarly, classical AAD clades—namely, (i) the cnidaria-algae clade; (ii) the clade encompassing cnidian, lophotrochozoan, and echinoderm representatives; and the SNAD4 clade—display length variations in loops 1 and 7. The arthropod clade shows length variations in loops 1 and 3. Loop 5, which hosts the cytotoxicity chelating the catalytic Zn$$^{2+}$$, is usually short in most clades but shows inserts between the two cysteine residues in the cnidaria-algae, nematode, and SNAD3 clades.

Thus, both low entropy values at target-interacting positions and low loop-length variability suggest that AID, APOBEC2, and APOBEC4 retain relatively conserved target specificities (Figs. 3E and 4). In contrast, most other AADs show concomitant variability in the shared target-recognition residues and inserts in the loops predicted to be close to the bound substrate (Figs. 3E and 4). Hence, it is possible that both variability at positions targeting the substrate and variable lengths of adjacent loops underlie the evolution of new target sequence specificities; in this way, the enzymes would counter emerging resistance of viral or parasite targets to deamination as a result of mutations in the target sequences. A second, but not mutually exclusive, possibility is that loop-length variation is associated with altered stability or evolution of additional interfaces for protein oligomerization. For instance, loop 3 of TadA and APOBEC3, respectively, is involved in dimerization and oligomerization. Oligomerization of APOBEC3 (e.g., APOBEC3G) has been shown to be essential for restriction of HIV-1 by facilitating its binding to the viral transcript strand to block the reverse transcriptase from catalyzing DNA elongation (49, 51). Thus, loop-length variations might also reflect accessory adaptations that extend or modulate the deaminase activity in biological conflicts with viruses and parasites.

**Divergence Arising from Potential Selective Pressures Imposed by Viral Inhibitors.** The HIV-1 Vif protein acts as a counter defense against APOBEC3s by binding to either of their tandem (APOBEC3G, APOBEC3F, and APOBEC3D) or solo (APOBEC3C/3H) deaminase domains; this triggers ubiquitin-mediated degradation by recruiting an E6/E3−CUL5−Rbx2 E3 ubiquitin-ligase complex (21). Mutagenesis suggests that Vif contacts APOBEC3s either through residues in helices 2, 3, and 4 and loop 4 (APOBEC3C, APOBEC3F), or through residues at the C-terminal part of loop 7 (APOBEC3H) (for review see ref. 52). A consensus emerging from several studies (52) implicates 13 residues commonly across APOBEC3s (L255, F258, C259, I262, S264, Y269, E289, F290, H294, D311, T312, D313, and E324 in APOBEC3F; PDB: 4j4j). However, several of these residues are not solvent-exposed, suggesting that they are only indirectly involved in transmitting conformational changes upon Vif binding. We asked whether the entropy of solvent-accessible positions in the above set of residues might reveal any general principles for interactions of AADs and viral inhibitors. Five of the above 13 positions, two in helix 2 (PDB: 4j4j; C259, E289), one in helix 4 (PDB: 4j4j; E324), and two in loop 7 (PDB: 4j4j; T312, D313), are both solvent-exposed and high in entropy (2.6–3.5; Fig. 3B). Investigation of neighboring solvent-exposed residues provided additional high-entropy positions in loop 4 and helix 3 (PDB: 4j4j; P265, A292, R293). Collectively, these positions locate to two distinct surfaces of the APOBEC3 structure (Fig. 3B): surface 1 includes residues at the C terminus of helix 2, the middle of loop 4, and the C terminus of helix 3; surface 2 is made up of residues from the C terminus of loop 7 and the N terminus of helix 5. Only surface 2 shows a small overlap with the DNA-binding interface (Fig. 3B). The additional parts of surface 2 are equivalent to the dimeric interface of TadA (51) or the oligomeric interface of APOBEC2 (53), respectively. The overlap of the binding site with a dimerization/oligomerization surface suggests that, in addition to recruiting a ubiquitin ligase complex, Vif might also interfere with oligomerization of APOBEC3s. Analysis of these regions in other AADs shows that, except for AID, APOBEC2, APOBEC4, NAD2, and the nematode AADs, the remainder show high positional entropy. Such a changing profile of exposed residues within and between clades suggests that most of the rapidly diversifying AADs are probably targeted by viral inhibitory factors analogous to Vif.

**Tandem Duplications, Gene Loss, Alternative Splicing, and Potential Intergenic Recombination in AADs.** As previously reported for the

---

**Fig. 4.** Length diversity of substrate-binding loops in AADs. AAD clades are listed on the y axis. For each clade, the range of the loop lengths in amino acids (x axis) is shown. The sequences corresponding to a loop of a particular length are shown as a circle centered at that length, with radius scaled by number of sequences normalized for a clade. The coloring scheme is a rainbow spectrum ranging from red (small number of sequences) to violet (large number of sequences), with the vertical blue lines representing the median loop length.
APOBECs (32), we noted that at least a subset of the paralogous AADs form an LSE cluster in the same genomic region. This was observed in several distant species, such as the dictyostelids *D. fasciculatum* and *T. lacteum*, the brachiopod *Lingula anatina*, sea urchins, the cnidian *Nematostella vectensis*, and lampreys (see companion paper). Like the intraspecific polymorphism of APOBEC3s with respect to paralog number (54), we found that the number of lamprey CDA1-like paralogs varies, and that they can be entirely lost even in individuals of the same species (see companion paper). To explore if such interindividual variations might be more generally observed, we performed whole-genome shotgun sequencing of the sea urchin *Strongylocentrotus purpuratus*, *Lingula*, and the oyster *Crassostrea gigas*, and assembled AAD sequences from these reads. We then compared these sequences with those from the individuals whose sequence is deposited in the GenBank database. In each case, we observed that, while the complement of AAD genes encoded by the genome was comparable, there were individual-specific differences (SI Appendix, Fig. S23), with some paralogous copies found only in one individual or the other (SI Appendix, Fig. S23). In *Nematostella*, we could identify an AAD pseudogene in the tandem gene array (SI Appendix, Fig. S23), indicating that paralog number variation could proceed via such pseudogenization events coupled with duplications in tandem gene arrays.

In certain metazoans, paralogous copies from a given species revealed the presence of a substantial segment of completely identical sequence that might be shared by one set of AAD paralogs, followed by a nonidentical segment distinguishing these paralogs. However, this nonidentical segment was found to be completely identical with the corresponding segment in another set of paralogs from the same organism. For example, the eight SNAD3 paralogs obtained from the assembled transcriptome of the emerald rockcod *Trematomus bernacchii* (SI Appendix, Fig. S24) were sequenced such that they all could be visualized as being constructed from different combinations of a small set of distinct segments: three N-terminal segments (nt1, nt2, nt3), two central segments (c1, c2), a single globally conserved segment (s1), and two distinct C-terminal segments (ct1, ct2) (SI Appendix, Fig. S24). A similar mosaic sequence pattern was also observed in the four *Nematostella* paralogs of the cnidaria-algae-APOBEC4 clade, in nine *Anthopleura* AADs, and in APOBEC1 paralogs from the turtle *Malaclemys terrapin* (SI Appendix, Fig. S24), which we recovered from transcriptome assemblies (55). This pattern is unlikley to emerge from a simple divergence of paralogous copies. In the case of *Trematomus* SNAD3 and the paralogs of echinoderm AADs, genomes are unavailable; hence, the exact basis of this mosaic pattern in paralogs obtained from the assembled transcriptome is unclear. In *Malaclemys* APOBEC1, all of the mosaicism was restricted to the C-terminal region of the paralogs. In the lamprey AADs, we detected a complex pattern of alternative splicing, giving rise to proteins with C-terminal diversity (see companion paper), suggesting that a similar process could be active in generating the APOBEC1 isoforms in *Malaclemys*. In *Nematostella*, when we compared the structure of the AAD paralogs deduced from the mRNA sequences in the assembled transcriptome with the corresponding genes in the genome, we found that, barring one paralog, the remainder did not have identical genomic counterparts. Rather, they appeared to be mosaic, with different segments corresponding to different genes (SI Appendix, Fig. S24).

Given that most of the coding frames of these genes derive from a single exon, barring a conflation in transcriptome assembly, this mosaic pattern of mRNAs possibly results from recombination between the different genomic copies. Since the transcriptomes were derived from adult somatic tissues (55), this would imply that such a recombination between paralogous genes in the genome occurs in somatic tissues to generate recombinant versions. Alternatively, it might result from ongoing recombination between the genomic copies in different individuals, given that the individuals from which the transcriptome and the genome sequences were derived are not the same.

**Major Evolutionary Trends in AADs and Their Functional Implications.** This study greatly extends the phylectic spread of the AADs in metazoaa (Fig. 5). Entropy analyses indicate high sequence variability of regions and residues involved in target motif termination, mutagenic activity, because it might be catalytically inactive versions of AADs. This is observed in some members of the APOBEC3 clade, AADs from *Ciona intestinalis*, and lophotrochozoan-echinoderm metal-binding clade AADs in the sea urchin *Evechinus chloroticus*. As proposed for inactive APOBEC3 domains (31, 32, 52), these inactive copies might function either as decoys to bind viral inhibitors or merely bind invasive nucleic acids without mutating them.

We interpret LSEs and rapid sequence and/or structural diversification, along with the frequent gene losses (Fig. 5), as signs of an ongoing arms race with pathogenic nucleic acids. By analogy to known AADs, we suggest that DNA viruses, retroviruses, and transposons are also targets for some of the AAD clades discovered herein. Many vertebrate AADs are expressed in lymphocytes, and this correlates with the remarkable lymphotropy of several retroviruses and DNA viruses (56, 57). Based on this observation, we propose that the biological conflict between AADs and viruses took shape in the lymphocyte or its invertebrate immunocyte counterpart. This provided a stepping stone for a second phase, during which “institutionalized” mutator variant(s) of the AADs, such as AID in gnathostomes and invertebrate immunocytes, proffered some intriguing aspects of the evolution of animal immunity. In a first step, viruses might have evolved to specifically infect immunocytes or lymphocytes, for this provides them with a host cell of astounding proliferative capacity. This also offered the viruses an opportunity to interfere with host innate or adaptive immune responses that relied on these cell types. In response, the host might have evolved mechanisms to deploy AADs in immune cells to neutralize lymphotropic viruses by mutagenic inactivation. Once this coevolutionary process was set in motion, it possibly provided the stepping stone for a second phase, during which “institutionalized” mutator variant(s) of the AADs, such as AID in gnathostomes and independently a subset of the CDAs in lampreys (20), were recruited. These facilitated the emergence of a fundamentally new aspect of adaptive immunity, namely somatic diversification of antigen-receptor genes by direct mutagenesis or by triggering DNA recombination.

In snails such as * Biomphalaria*, somatic diversification of the polymorphic plasma lectins, the fibrinogen-related proteins (FREPs), is proposed to be a component of an anticipatory immune system (58–60). Analogous to vertebrate immunoglobulins, FREPs recognize antigenic variations in snail parasites such as trematodes and exhibit signatures of somatic mutations (60). We identified a subclade of AADs within the lophotrochozoan-echinoderm-metal-binding clade, which is specifically conserved across gastropods, including * Biomphalaria* (APKA01034104.1).
Figure 5. Evolutionary reconstruction and origins of various AAD families. The provenance/distribution of AAD clades is superimposed on a simplified eukaryotic tree. Presence of AADs (green), LSEs of AADs (red circles with outer glow), and absence/potential loss of AADs (yellow markers/markers with cross). Tunicate inactive versions (gray circle); dotted lines indicate alternative lateral transfer scenarios from bacteria to stems of metazoa and dictyostelids or to stems of opisthokonta-amoebozoa with loss in fungi (**).
35. Petersen-Mahrt SK, Neuberger MS (2003) In vitro deamination of cytosine to uracil in DNA. J Mol Biol 335:635-656.

36. Almgren R, Maley F, Maley GF, Maccoll R, Van Roey P (2004) Three-dimensional structural model of T4-bacteriophage 2-deoxycytidine deaminase. J Biol Chem 279:35479-35485.

37. Kumasaka T, et al. (2007) Crystal structures of bacterial S deaminases (SD): Implications for dynamics properties of catalytic zinc. J Biol Chem 282:37103-37111.

38. Sternmark P, Moche M, Gurmu D, Nordlund P (2007) The crystal structure of the bifunctional deaminase/reductase RbdB of the riboswitch biosynthetic pathway in Escherichia coli: Implications for the reductive mechanism. J Mol Biol 373:68-84.

39. Wolf J, Gerber AP, Keller W (2002) tADa, an essential tRNA-specific adenine deaminase from Escherichia coli. EMBO J 21:3841-3851.

40. Nishikura K (2010) Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem 79:321-49.

41. Salone V, et al. (2007) A hypothesis on the identification of the editing enzyme in animals. EMBO J 26:2336-2347.

42. Shelnutt JA (2015) RNA editing in plants. Machinery and flexibility of the recognition. Biochim Biophys Acta 1847:779-785.

43. Conticello SV (2008) The AID/APOBEC family of nucleic acid mutators. Genome Biol 9:229.

44. Saltz JD, Bennett RP, Smith HC (2016) The APOBEC protein family: Unified by structure, divergent in function. Trends Biochem Sci 41:578-594.

45. Luo M, Schramm VL (2008) Transition state structure of E. coli tRNA-specific adenine deaminase. J Am Chem Soc 130:2649-2655.

46. Gerber A, Grojean H, Melcher T, Keller W (1998) Tad1p, a yeast tRNA-specific adenine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. EMBO J 17:4780-4789.

47. Maas S, Melchers F, Rich A (1999) Identification and characterization of a human adenosine deaminase related to the ADAR family of pre-mRNA editing enzymes. Proc Natl Acad Sci USA 96:8895-8900.

48. Nishikura K (2016) A-to-I editing of coding and non-coding RNAs by ADARs. Mol Cell 63:956-969.

49. Harris RS, Siddamant MT (2004) Retroviral restriction by APOBEC3 proteins. Nat Rev Immunol 4:867-877.

50. Iyer LM, Zhang D, Rogozin IB, Altschul SF (2005) Crystal structure of the editing enzyme in human B-cell non-Hodgkin lymphomas. Blood 101:3574-3580.

51. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Bioinformatics 21:368-370.

52. Aydin H, Taylor MW, Lee JE (2014) Structure-guided analysis of the human APOBEC3-B Editing Complex. Structure 22:668-684.

53. Prochnow C, Bransteitter R, Klein MG, Goodman MF, Chen XS (2007) The APOBEC-3D crystal structure and functional implications for the deamination AID. Nature 445:447-451.

54. Münk C, Willemsen A, Bravo IG (2012) An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol 12:71.

55. Babonis LS, Martindale MQ, Ryan JF (2016) Do novel genes drive morphological variation? An investigation of the nematodes in the sea anemone Nemastomella vectensis. BMC Biol 16:114.

56. Reisfeld EW, et al. (2010) Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: Implications for HIV-1 restriction. Nucleic Acids Res 38:4274-4284.

57. Jones J, et al. (2003) Expression of activation-induced cytidine deaminase in human B-cell non-Hodgkin lymphomas. Blood 101:3574-3580.

58. Adema CM (2015) Fibrinogen-related proteins (FREPs) in mollusks. Results Probl Cell Diff 57:111-129.

59. Léonard PM, Adema CM, Zhang SM, Loker ES (2001) Structure of two FREP genes that encode fibrinogen-related proteins (FREPs) from an oceanic mollusk, T. vectensis.

60. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein similarity detection and structure prediction. Nucleic Acids Res 25:3389-3402.

61. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 20:203-211.

62. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European molecular biology open software suite. Trends Genet 16:276-277.

63. Metzger MJ, et al. (2016) Widespread transmission of independent cancer lineages within multiple bivalve species. Nature 534:795-799.

64. Nicolaus A, et al. (2018) Secreted heme peroxidase from Dictyostelium discoideum: Insights into catalytic structure, function, and biological role. J Biol Chem 293:1330-1345.

65. Mungpakdee S, et al. (2014) Massive gene transfer and extensive RNA editing of a bacterial monocistronic plasmid genome. Genome Biol 15:1428.

66. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

67. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205-211.

68. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European molecular biology open software suite. Trends Genet 16:276-277.

69. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244-W248.

70. Metzger MJ, et al. (2016) Widespread transmission of independent cancer lineages within multiple bivalve species. Nature 534:795-799.

71. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

72. Cooke C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36:W197-W201.

73. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—Approximately maximum-likelihood trees for large alignments. PLoS One 5:e10869.

74. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874.