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

Arunkumar Krishnan[a], Lakshminarayan M. Iyer[a], Stephen J. Holland[b], Thomas Boehm[b], and L. Aravind[a,1]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; and [b]Department 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, dictyosteliida, 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 APOBECs1–3, 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.

DNA/RNA editing | immunity | arms race | biological conflicts | retroelements

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 of bases in free nucleotides, such as the cytidine deaminases (CDD/CDA) (4), deoxycytidylate monophosphate deaminases (dCMP) (5), and guanine deaminase (GuaD) (6); (*ii*) engaged in the biosynthesis of modified nucleotide-derived compounds (e.g., blasticidin-S deaminase) (7) and the RibD deaminase domain (8); and (*iii*) active in in situ modifications of bases in nucleic acids, such as tRNA adenosine deaminases (Tad2/TadA and Tad1) (9), RNA-specific adenosine deaminase (ADAR) (10), RNA-editing cytidine 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 DNA 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_nC (where x is any amino acid and n is ≥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, blasticidin-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, dictyosteliid, and algal lineages. They appear to have emerged from an early transfer of an AAD from bacterial toxin 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.

Data deposition: The sequence reported in this paper has been deposited in the Sequence Read Archive at the NCBI (accession nos. SAMN08013505, SAMN08013506, and SAMN08013507).

[1] 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.

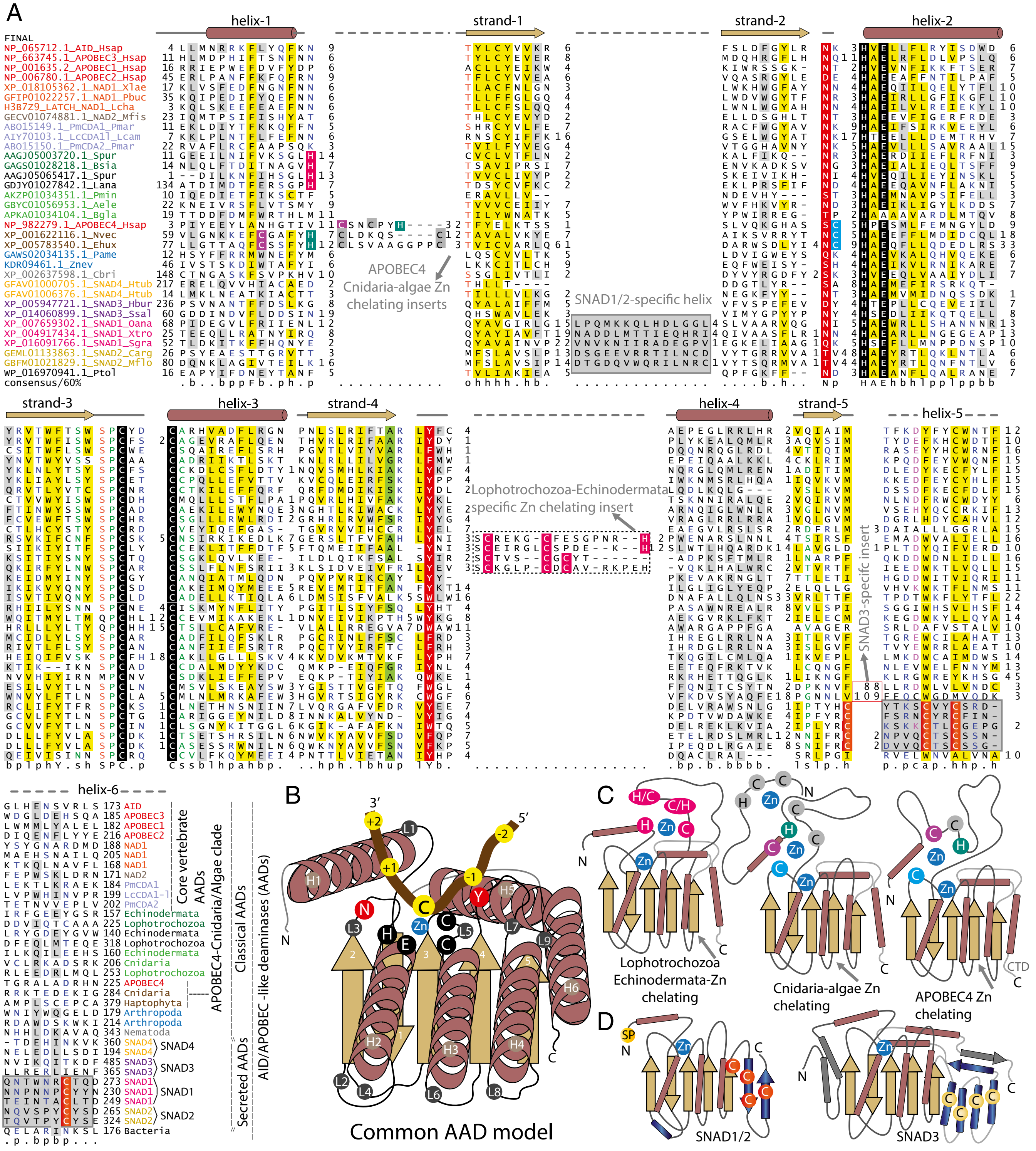

Fig. 1. Characterization of the AAD protein family. (*A*) MSAs of representatives of AAD clades labeled with accessions, clade names, and species abbreviations (full species names in *SI Appendix*, Fig. S1). Zn-chelating active site residues are highlighted in black, and known substrate-interacting residues 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 SNADs1–3: 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 agnathan (jawless vertebrate) lymphocytes (PmCDA1 and PmCDA2) 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, parvoviruses, papillomaviruses, hepatitis B virus, and various retroelements (31–33). APOBEC1 deaminates cytosine both in RNA (34) and ssDNA (35) and has roles in both mRNA editing and ssDNA mutagenesis as part of the defense against retroviruses and genomic 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 toxin 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-undiscovered 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 present evidence 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 cyclostome 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 and Transcriptome Shotgun Assembly (TSA) and Whole-Genome-Shotgun (WGS) contig databases using various known AADs as queries (see *Materials and Methods*). Profile searches were run using PSI-BLAST and JACKHMMER 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, urochordates, echinoderms, lophotrochozoans, arthropods, cnidarians, poriferans, dictyosteliids, and algae, and added homologs from species within taxa that were previously known to possess AADs (Fig. 1*A*; *SI Appendix*, Fig. S1; c.f., phyletic distribution in *SI Appendix*, Fig. S2).

We identified potential AADs in a flatworm (platyhelminthes), *Microphallus* sp., and its snail host, *Potamopyrgus antipodarum*; a green algae, *Elliptochloris marina*; an endosymbiont of sea anemone, *Anthopleura elegantissima*; and numerous dinoflagellates that are symbionts of corals. In these instances, the complete sequence identities between the host and symbiont/pathogen sequences, coupled with the results from phylogenetic tree analysis, show that the *Microphallus*, *Elliptochloris*, and the dinoflagellate AAD sequences are likely contaminations from their host genome sequences; hence, these sequences were removed before further analysis.

To understand the evolution of the AADs in light of the newly detected members, we built multiple sequence alignments (MSAs; Fig. 1*A* 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. 1*A* and 2*A*). 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 analysis further clarified the evolutionary trajectories and relationships for both the previously known AADs and the newly detected members (Fig. 2*A*). 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 chondrichthyans (sharks), actinopterygians (ray-finned fishes), and sarcopterygians (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 actinopterygians and amphibians (*SI Appendix*, Figs. S2–S4). Fourth, APOBEC3 appears 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, lophotrochozoans, and cnidarians, and APOBEC4 form multiple distinct clades that are out-groups to the above core vertebrate AAD clade (Fig. 2*A*). We refer to the clade that unites all these AADs with core vertebrate AADs as the classical AAD clade (Fig. 2*A*). Of these, APOBEC4 is recovered from tetrapods, sarcopterygians, and agnathans, but is frequently lost across actinopterygians (Fig. 2*A* and *SI Appendix*, Figs. S2 and S13). Based on conserved sequence features and a long, distinctive, shared metal-binding insert (Fig. 1*C*, *Right* panels, and *SI Appendix*, Fig. S14), 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 metazoa 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 possess an N-terminal signal peptide (Figs. 1 and 2*A* and *SI Appendix*, Figs. S15–S18). This feature suggests that these AADs might be secreted; accordingly, we named these the Secreted Novel AID/APOBEC-like Deaminases (SNADs). 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. 2*A*). We also detected secreted AADs in the dictyosteliid slime molds *Dictyostelium fasciculatum*, *Tieghemostelium lacteum*, and *Acytostelium*, which might either group with the SNADs or form a distinct basal clade of eukaryote-specific AADs (Fig. 2*A* and *SI Appendix*, Fig. S19) We 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. 2*A*). 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 2*A*): (*i*) a clade found in cnidarians, lophotrochozoans, echinoderms, and tunicates with a deaminase domain most closely related to the core vertebrate AADs

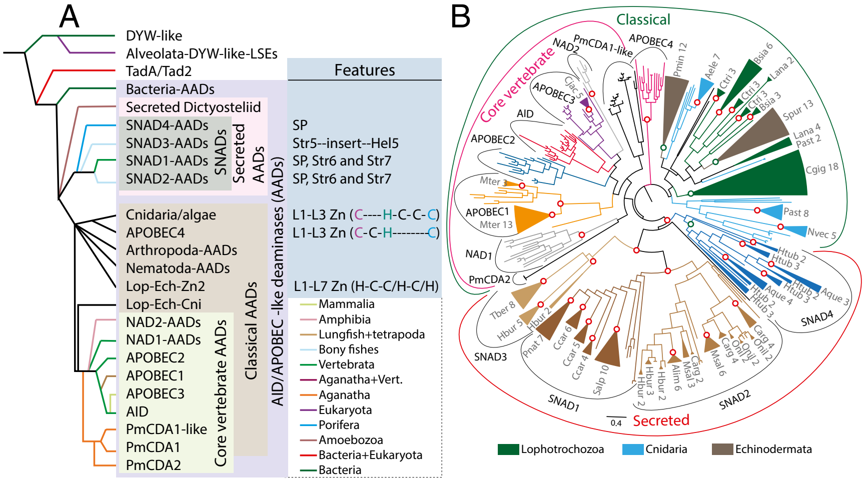

Fig. 2. Phylogenetic relationships among AADs. (*A*) Higher order relationships of deaminases with a focus on AADs. Clade-specific features are on the *Right*. Zn-chelating residues are shown within parentheses and homologous residues shared by members of the cnidaria-algae-APOBEC4 clade are colored. (*B*) Phylogenetic tree illustrating LSEs of AADs in various metazoan lineages. Clades entirely composed of monospecific representatives are collapsed and labeled with species abbreviations and number of sequences in the LSE. Nodes supported by bootstrap values >80% and >90% are marked with green-outlined and red-outlined circles, respectively (also see *SI Appendix*, Fig. S21). Hel, helix; L, loop; SP, signal peptide; Str, strand.

(*SI Appendix*, Fig. S9); (*ii*) a clade found in lophotrochozoans and echinoderms with a metal-chelating 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. 2*A* and *SI Appendix*, Figs. S11 and S12).

The invertebrate classical AADs are dominated by LSEs, ranging from 2 to 22 paralogs per taxon (Fig. 2*B* 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 *Biomphalaria glabrata* appears to possess only a single representative. In general, several species of lophotrochozoa 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, SNADs1–4. SNAD1, SNAD2. and SNAD3 clades are specific to vertebrates, whereas SNAD4 is found only in sponges (Fig. 2 *A* and *B* and *SI Appendix*, Figs. S2 and S15–S18). In vertebrates, SNAD1 shows the widest phyletic spread, being present in amphibians, reptiles, the monotreme mammal the platypus, the coelacanth, and actinopterygians, whereas SNAD2 and SNAD3 are only found in actinopterygians. Hence, in vertebrates, SNAD1 is likely the ancestral clade that goes back to at least the stem euteleostomian, while SNAD2 and SNAD3 were derived from it as actinopterygian-specific expansions. SNADs show repeated LSEs in certain actinopterygians and turtles, while SNAD4 members are found as LSEs even in individual sponge species (Fig. 2*B* and *SI Appendix*, Figs. S2 and S21). The SNADs have accreted several unique structural features around the core deaminase fold: SNAD1 and SNAD2 share a helical insert between strands 1 and 2 (Fig. 1*D* and *SI Appendix*, Figs. S15 and S16). In place of the usual helices 5 and 6 of the deaminase core domain, they contain a neomorphic strand 6 and 7, which we predict to form a β-hairpin packing with the core strand 5 (Fig. 1 *A* and *D* and *SI Appendix*, Figs. S15 and S16). Furthermore, SNAD1 and SNAD2 possess a unique cluster of four cysteine residues with the first cysteine at the end of strand 5, a CxxC motif in the middle of strand 6, and the last cysteine in the middle of strand 7, which might form disulfide linkages to stabilize the secreted protein (Fig. 1 *A* and *D* and *SI Appendix*, Figs. S15 and S16). SNAD3 lacks the above insert, has an N-terminal helical extension in lieu of the signal peptide, and a distinct helical insert in the loop-9 region (Fig. 1*D* and *SI Appendix*, Fig. S17). Additionally, these clades show unique lineage-specific sequence synapomorphies (*SI Appendix*, Table S1).

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 dictyosteliids are unique among the eukaryotic deaminases in being secreted. The dictyosteliid AADs possess signal peptides like SNAD1, SNAD2, and SNAD4 and likewise show independent lineage-specific expansions (6–7 copies) in *D. fasciculatum* and *T. lacteum* (*SI Appendix*, Fig. S19). They are distinguished by a distinct C-terminal domain with four cysteines that are likely to form disulfide bonds (*SI Appendix*, Fig. S19). However, their extreme divergence prevents us from establishing a specific relationship between them and the metazoan SNADs. Together, these secreted versions resemble the ancestral AADs from bacterial polymorphic toxins (22, 39), which were the likely precursors of eukaryotic AADs. Another functional analogy is offered by the nucleic-acid–targeting CR (Crinkler-RHS–type) effectors, which are secreted by various eukaryotes and delivered into recipient cells (40). The SNADs and dictyosteliid AADs might be similarly 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. Hence, they might be deployed against pathogens specifically affecting organisms with 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 Zn^{2+} ion (Figs. 1*B* and 3*A*). In both TadA [Protein Data Bank (PDB): 2b3j] and APOBEC3A (PDB: 5sww), 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. 3*A*). 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. 3*A*): (*i*) the active site histidine (PDB:

5sww: H70; PDB: 2b3j: H53) establishes π-π interactions; (*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. 1*A* and 3 *A* 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 the loop 7 of TadA result in a larger substrate-binding pocket accommodating an adenine (42). Additional residues in loops 1, 3, 5, and 7 and strand 4 contact the target base mainly via polar interactions with the polypeptide backbone, or interact with the sugar-phosphate backbone of the nucleic acid (Fig. 3*A*). One of these is the position at the end of strand 4, which is a strongly conserved basic residue (R128; PDB: 5ssw) 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 positive selection 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).
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 is indicative of 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.

Diversification of Regions Recognizing the Target Sequence in AADs. We then investigated local variations in positional entropy at sites potentially associated with substrate interaction, aided by published cocrystal structures of deaminases with their substrates. Comparison of the substrate-bound TadA and APOBEC3A structures (42, 43) shows that they make extensive contacts extending up to two bases each on the 5′ and 3′ sides of the target base, implying a 5-nucleotide-long recognition sequence. These ancillary contacting residues emerge from the same loops (1, 3, 5, 7) that also contact the target base (Fig. 3*A*). A considerable diversity of interactions is observed between the polypeptide side chains or backbone with the bases and the sugar-phosphate backbone of the nucleic acid. Of these interactions in APOBEC3A (PDB: 5sww), the side chains of D131 and Y132 in loop 7 and W98 in loop 5 make direct polar or van der Waals contacts with the −1 base of the target DNA (43). Similarly, side chains of H29 and K30 residues in loop 1 of APOBEC3A respectively make π-π stacking and cation-π interactions with the +1 base, whereas R28 in loop 1 makes a cation-π contact with the −2 base. Characterized AADs show different target sequence specificities, often with some degeneracy. AID prefers a WR$\underline{C}$ (W: A/T, R: A/G) motif, whereas APOBEC3A prefers thymine at the −1 and cytosine at the −2 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. 3*E*). In contrast, APOBEC1, APOBEC3, most invertebrate classical AAD clades, and the SNADs show high sequence entropy at these positions (Fig. 3 *D* 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 include polarity shift (polar vs. hydrophobic), charge inversion (positive vs. negative), or gain/loss of charge (charge vs. neutral) (Fig. 3*E*), highlighting the potential positive selection on these proteins.

Substrate-binding loops might also display dramatic length variations (Fig. 4). We consistently observed a tendency for convergent emergence of long inserts in different loops stabilized by metal chelating residues or disulfide bonds that are defining features of particular clades (Fig. 1 *C* and *D* and *SI Appendix*, Table S1),

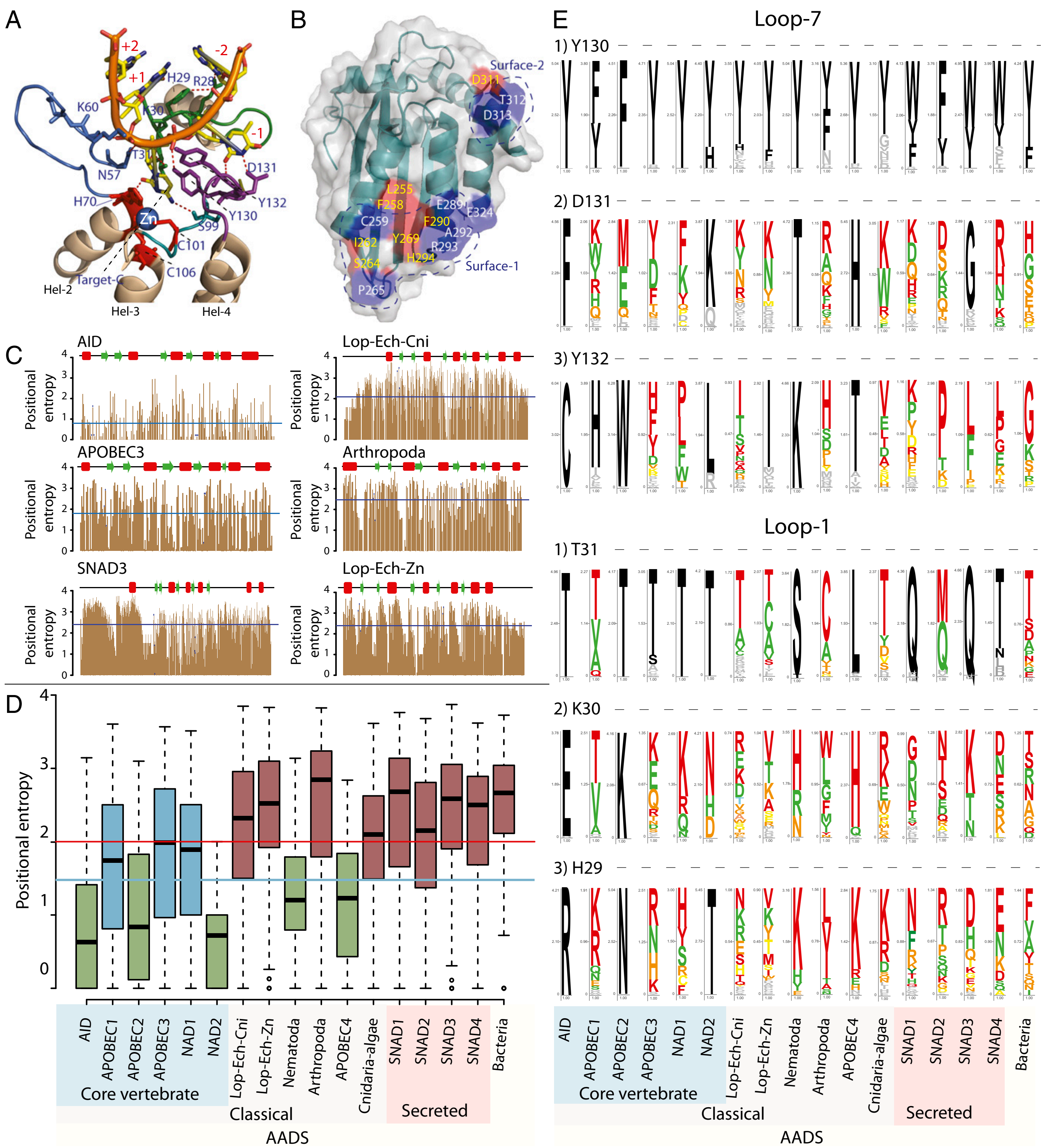

Fig. 3. (*A*) Illustration of human APOBEC3A bound to ssDNA substrate (PDB ID code 5ssw) showing key APOBEC3A residues involved in Zn chelation and interactions with the target and neighboring bases (labeled −2, −1, +1, and +2). (*B*) Surface view of APOBEC3F (PDB: 4j4j) depicting Vif1-binding residues 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 cnidarian 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 dictyosteliid 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 dictyosteliid 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 ≥42 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 cnidarian, 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 cysteines 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. 3*E* 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. 3*E* 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 template 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 role of 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 EloB/C-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 (APOBEC 3C, APOBEC 3F), 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. 3*B*). 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. 3*B*): 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. 3*B*). 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

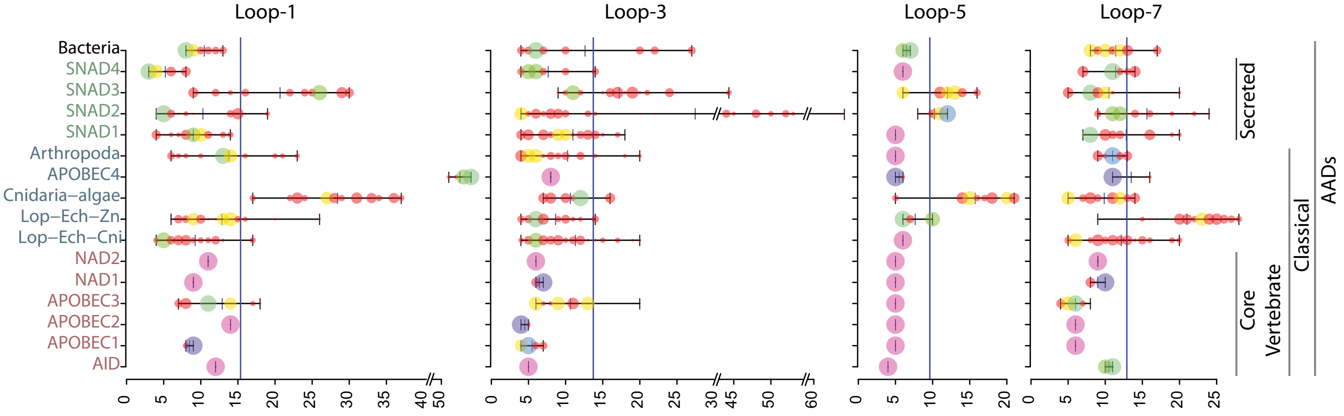

Fig. 4. Length diversity of substrate-binding loops in AADs. AAD clades are listed on the *y* axis. For each clade, the range of the loop lengths in amino acids (*x* axis) is shown. The sequences corresponding to a loop of a particular length are shown as a circle centered at that length, with radius 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.

APOBEC3s (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 dictyosteliids *D. fasciculatum* and *T. lacteum*, the brachiopod *Lingula anatina*, sea urchins, the cnidarian *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) suggested that they all could be visualized as being constituted 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 unlikely to emerge from a simple divergence of paralogous copies. In the case of *Trematomus* SNAD3 and the *Anthopleura* AADs, the 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 the 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 phyletic spread of the AADs in metazoa (Fig. 5). Entropy analyses indicate high sequence variability of regions and residues involved in target motif determination, multimerization, or where potential inhibitory factors analogous to Vif may bind. This is observed for several AAD clades (e.g., APOBEC1, APOBEC3), almost all invertebrate clades (except for the nematode AADs), the SNADs, and dictyosteliid AADs, suggesting that these genes are under positive selection. The evolutionary trends among the AAD clades can be subsumed under the following four categories (Fig. 5):

i) Clades that typically contain one well-conserved copy per species. These are likely involved in conserved functions like AID; in turn, this suggests that comparable conserved editing roles are likely performed by APOBEC2 and APOBEC4.

ii) Fast-evolving clades characterized by numerous paralogous genes within each species. Here, further variability via tandem duplications and gene loss, alternative splicing, or intergenic recombination might also be at work, e.g., the APOBEC3s, SNADs, dictyosteliid AADs, several invertebrate AADs, and CDA1-like genes in lampreys.

iii) Rapidly evolving clades typically present in relatively low copy number, such as those from the insect lineages polyneoptera, mecoptera, and zygentoma.

iv) Repeated emergence of 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 proposal is attractive, because it illuminates 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)

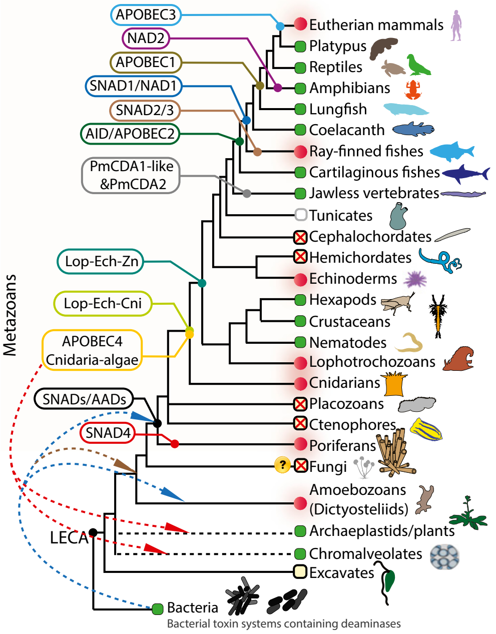

Fig. 5. Evolutionary reconstruction and origins of various AAD families. The provenance/distribution of AAD clades is superimposed on a simplified eukaryotic tree. Presence of AADs (green), LSEs of AADs (red circles with outer glow), and absence/potential loss of AADs (yellow markers/markers with cross). Tunicate inactive versions (gray circle); dotted lines indicate alternative lateral transfer scenarios from bacteria to stems of metazoa and dictyosteliids or to stems of ophisthokonta-amoebozoa with loss in fungi ("?").

and *Aplysia californica* (GBBE01059801.1). These AADs are distinguished from the rest of the clade by a loss of the conserved glutamate in the HxE motif but possess a compensatory conserved E after the second C of the Cx_nC motif, which might similarly chelate the catalytic Zn^{2+} (*SI Appendix*, Fig. S10). Like AID in gnathostomes, a single representative of this group is conserved across diverse gastropods irrespective of whether the organisms possess other LSEs of AADs. We propose that these molluscan AADs might play a role in the mutagenic variability of FREPs and that certain other invertebrate classical AADs identified herein possibly also perform such mutator roles.

The discovery of SNADs hints at the possibility of unexplored contexts for AAD function. The prediction that they are secreted suggests that SNADs function either in the endoplasmic reticulum or in the extracellular space. Several RNA and DNA viruses replicate in specialized viral factories arising from the endoplasmic reticulum (61); moreover, several other viruses exploit the vesicular trafficking system during uptake and release. Thus, SNADs might specialize in targeting viruses in vesicles as proposed for certain APOBEC3s (62). As noted above, vertebrate SNADs are confined to animals with poikilothermy or low body temperature. Hence, they might be secreted to specifically target pathogens affecting such organisms (e.g., certain fungi) (63). They might also be deployed as mutators directed against other extracellular parasites or even pathogenic cells (e.g., transmissible tumors) (64). Interestingly, in the slime mold *T. lacteum*, we identified an ∼16.6-kb potential immunity-related locus with six tandem genes coding for secreted proteins (GenBank accession no. LODT01000051.1). In addition to three of the six secreted AADs from this organism, the locus also encodes two paralogous MAC-Perforin domain proteins (also showing gene expansions in slime molds) and a heme peroxidase, which has previously been implicated in antibacterial defense in slime molds (65). This supports the idea that secreted AADs might play a role in immunity more widely in eukaryotes.

Various aspects of deaminase diversification described above might also play out more widely across other distantly related deaminases. In this study, we noticed a similar LSE of the DYW clade of deaminases in dinoflagellates of diverse lineages, including *Gonyaulacales*, *Suessiales*, *Syndiniales*, and *Noctilucales* (*SI Appendix*, Fig. S25). Genomes of individual species might contain anywhere from 1 to 39 paralogous DYW deaminases. Based on the widespread presence of RNA editing in chloroplasts of *Symbiodinium minutum* (66), it is possible that the DYW deaminases are involved in this process, perhaps in a gene-specific way as proposed for the plant chloroplasts. However, it is also possible that the expansions reflect an adaptation to the conflict with viruses or transposons, similar to what we had earlier proposed for comparable expansions of another clade of deaminases in fungi (22).

In conclusion, this report provides the basis for further investigations of mutagenic deaminases in as yet poorly studied biological contexts. We suggest that some of these enzymes may also serve as reagents for biotechnological purposes.

Materials and Methods

Sequence Searches. Sequence searches were conducted using the PSI-BLAST (67) and JACKHMMER (68) algorithms against the nonredundant (nr) database at the NCBI and all ENSEMBL database proteomes. Transcriptome Shotgun Assembly (TSA) and WGS contig databases were queried using TBLASTN. Open reading frames (ORFs) were predicted using the getorf program from EMBOSS 6.5.7 software package (69). Profile-Profile searches were run using the HHpred (70) program either against the PDB or Pfam databases. MSAs were built using the KALIGN, MAFFT, and GISMO programs (71). Secondary structures were predicted with the JPred program (72).

WGS Libraries. Individual *C. gigas* (procured from the Isle of Sylt, Germany) and *L. anatina* (caught in Nha Trang Bay, Vietnam) and *S. purpuratus* (caught off the Californian coast) specimens were used to extract genomic DNA (total body for *C. gigas* and *L. anatina*; coelomocytes for *S. purpuratus*). Libraries were made from a minimum of 1 μg of total genomic DNA and were sequenced at 2 × 250 bp paired end to a depth of 150 million reads using HiSeq2500 Rapid Mode (Illumina). Adapter sequences were trimmed off the read ends, and read files were filtered through quality-control checks and formatted into nucleotide databases (*SI Appendix, SI Methods*). Sequences were deposited in the Sequence Read Archive at the NCBI with the following accession numbers: SAMN08013505 (*C. gigas*), SAMN08013506 (*S. purpuratus*), and SAMN08013507 (*L. anatina*).

Phylogenetic Trees. Phylogenetic relationships were derived using an approximate maximum likelihood (ML) method as implemented in the FastTree program (73) and the full ML method implemented in MEGA7 (74). To increase the accuracy of topology in FastTree, we increased the number of rounds of minimum-evolution subtree-prune-regraft (SPR) moves to 4 (-spr 4) as well as utilized the options -mlacc and -slownni to make the ML nearest neighbor interchanges (NNIs) more exhaustive.

Entropy Analysis. Position-wise Shannon entropy (H) was computed using a custom script written in the R language using the equation

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

where M is the number of amino acid types and P is the fraction of residues of amino acid type i. The Shannon entropy for any given position in the MSA ranges from 0 (absolutely conserved one amino acid at that position) to 4.32 (all 20 amino acid residues equally represented at that position).

Data Visualization. Structures were visualized and compared using the PyMOL program (https://pymol.org/2/). R language scripts were used for analysis of data and generation of entropy and loop-length divergence plots.

ACKNOWLEDGMENTS. We thank Christian Buschbaum, Elena Temereva, and Robert A. Cameron for help with obtaining animal specimens. A.K., L.M.I., and L.A. are supported by the intramural funds of the National Library of Medicine, NIH. S.J.H. and T.B. are supported by the Max Planck Society and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013; ERC Grant Agreement 323126).

1. Smith HC, Gott JM, Hanson MR (1997) A guide to RNA editing. *RNA* 3:1105–1123.
2. Gott JM, Emeson RB (2000) Functions and mechanisms of RNA editing. *Annu Rev Genet* 34:499–531.
3. Hamilton CE, Papavasiliou FN, Rosenberg BR (2010) Diverse functions for DNA and RNA editing in the immune system. *RNA Biol* 7:220–228.
4. Betts L, Xiang S, Short SA, Wolfenden R, Carter CW, Jr (1994) Cytidine deaminase. The 2.3 A crystal structure of an enzyme: Transition-state analog complex. *J Mol Biol* 235: 635–656.
5. Almog R, Maley F, Maley GF, Maccoll R, Van Roey P (2004) Three-dimensional structure of the R115E mutant of T4-bacteriophage 2′-deoxycytidylate deaminase. *Biochemistry* 43:13715–13723.
6. Liaw SH, Chang YJ, Lai CT, Chang HC, Chang GG (2004) Crystal structure of Bacillus subtilis guanine deaminase: The first domain-swapped structure in the cytidine deaminase superfamily. *J Biol Chem* 279:35479–35485.
7. Kumasaka T, et al. (2007) Crystal structures of blasticidin S deaminase (BSD): Implications for dynamic properties of catalytic zinc. *J Biol Chem* 282:37103–37111.
8. Stenmark P, Moche M, Gurmu D, Nordlund P (2007) The crystal structure of the bifunctional deaminase/reductase RibD of the riboflavin biosynthetic pathway in Escherichia coli: Implications for the reductive mechanism. *J Mol Biol* 373:48–64.
9. Wolf J, Gerber AP, Keller W (2002) tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *EMBO J* 21:3841–3851.
10. Nishikura K (2010) Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* 79:321–349.
11. Salone V, et al. (2007) A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Lett* 581:4132–4138.
12. Shikanai T (2015) RNA editing in plants: Machinery and flexibility of site recognition. *Biochim Biophys Acta* 1847:779–785.
13. Conticello SG (2008) The AID/APOBEC family of nucleic acid mutators. *Genome Biol* 9: 229.
14. Salter JD, Bennett RP, Smith HC (2016) The APOBEC protein family: United by structure, divergent in function. *Trends Biochem Sci* 41:578–594.
15. Luo M, Schramm VL (2008) Transition state structure of E. coli tRNA-specific adenosine deaminase. *J Am Chem Soc* 130:2649–2655.
16. Gerber A, Grosjean H, Melcher T, Keller W (1998) Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. *EMBO J* 17:4780–4789.
17. Maas S, Gerber AP, Rich A (1999) Identification and characterization of a human tRNA-specific adenosine deaminase related to the ADAR family of pre-mRNA editing enzymes. *Proc Natl Acad Sci USA* 96:8895–8900.
18. Nishikura K (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* 17:83–96.
19. Bass BL, Weintraub H (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* 55:1089–1098.
20. Rogozin IB, et al. (2007) Evolution and diversification of lamprey antigen receptors: Evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nat Immunol* 8: 647–656.
21. Harris RS, Liddament MT (2004) Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* 4:868–877.
22. Iyer LM, Zhang D, Rogozin IB, Aravind L (2011) Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Res* 39:9473–9497.
23. Muramatsu M, et al. (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102:553–563.
24. Kinoshita K, Honjo T (2001) Linking class-switch recombination with somatic hypermutation. *Nat Rev Mol Cell Biol* 2:493–503.
25. Moris A, Murray S, Cardinaud S (2014) AID and APOBECs span the gap between innate and adaptive immunity. *Front Microbiol* 5:534.
26. Litman GW, Rast JP, Fugmann SD (2010) The origins of vertebrate adaptive immunity. *Nat Rev Immunol* 10:543–553.
27. Boehm T, Swann JB (2014) Origin and evolution of adaptive immunity. *Annu Rev Anim Biosci* 2:259–283.
28. Harris RS, Sale JE, Petersen-Mahrt SK, Neuberger MS (2002) AID is essential for immunoglobulin V gene conversion in a cultured B cell line. *Curr Biol* 12:435–438.
29. Bastianello G, Arakawa H (2017) A double-strand break can trigger immunoglobulin gene conversion. *Nucleic Acids Res* 45:231–243.
30. Boehm T, et al. (2012) VLR-based adaptive immunity. *Annu Rev Immunol* 30:203–220.
31. Stavrou S, Ross SR (2015) APOBEC3 proteins in viral immunity. *J Immunol* 195: 4565–4570.
32. Refsland EW, Harris RS (2013) The APOBEC3 family of retroelement restriction factors. *Curr Top Microbiol Immunol* 371:1–27.
33. Duggal NK, Emerman M (2012) Evolutionary conflicts between viruses and restriction factors shape immunity. *Nat Rev Immunol* 12:687–695.
34. Fossat N, et al. (2014) C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. *EMBO Rep* 15:903–910.
35. Petersen-Mahrt SK, Neuberger MS (2003) In vitro deamination of cytosine to uracil in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1). *J Biol Chem* 278:19583–19586.
36. Ikeda T, et al. (2017) Opossum APOBEC1 is a DNA mutator with retrovirus and retroelement restriction activity. *Sci Rep* 7:46719.
37. Ikeda T, et al. (2008) The antiretroviral potency of APOBEC1 deaminase from small animal species. *Nucleic Acids Res* 36:6859–6871.
38. Conticello SG, Thomas CJ, Petersen-Mahrt SK, Neuberger MS (2005) Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol Biol Evol* 22: 367–377.
39. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L (2012) Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct* 7:18.
40. Zhang D, Burroughs AM, Vidal ND, Iyer LM, Aravind L (2016) Transposons to toxins: The provenance, architecture and diversification of a widespread class of eukaryotic effectors. *Nucleic Acids Res* 44:3513–3533.
41. Bergman A, Casadevall A (2010) Mammalian endothermy optimally restricts fungi and metabolic costs. *MBio* 1:e00212-10.
42. Losey HC, Ruthenburg AJ, Verdine GL (2006) Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA. *Nat Struct Mol Biol* 13: 153–159.
43. Shi K, et al. (2017) Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nat Struct Mol Biol* 24:131–139.
44. Gaudelli NM, et al. (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551:464–471.
45. Rubio MA, et al. (2007) An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc Natl Acad Sci USA* 104:7821–7826.
46. Powell C, Cornblath E, Goldman D (2015) Zinc-binding domain-dependent, deaminase-independent actions of apolipoprotein B mRNA-editing enzyme, catalytic polypeptide 2 (Apobec2), mediate its effect on zebrafish retina regeneration. *J Biol Chem* 290:6007.
47. Sato Y, et al. (2010) Deficiency in APOBEC2 leads to a shift in muscle fiber type, diminished body mass, and myopathy. *J Biol Chem* 285:7111–7118.
48. Pham P, Bransteitter R, Petruska J, Goodman MF (2003) Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* 424:103–107.
49. Shandilya SM, et al. (2010) Crystal structure of the APOBEC3G catalytic domain reveals potential oligomerization interfaces. *Structure* 18:28–38.
50. Marx A, Galilee M, Alian A (2015) Zinc enhancement of cytidine deaminase activity highlights a potential allosteric role of loop-3 in regulating APOBEC3 enzymes. *Sci Rep* 5:18191.
51. Kuratani M, et al. (2005) Crystal structure of tRNA adenosine deaminase (TadA) from Aquifex aeolicus. *J Biol Chem* 280:16002–16008.
52. Aydin H, Taylor MW, Lee JE (2014) Structure-guided analysis of the human APOBEC3-HIV restrictome. *Structure* 22:668–684.
53. Prochnow C, Bransteitter R, Klein MG, Goodman MF, Chen XS (2007) The APOBEC-2 crystal structure and functional implications for the deaminase 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 novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis. *BMC Evol Biol* 16:114.
56. Refsland 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. Greeve 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 Differ* 57:111–129.
59. Léonard PM, Adema CM, Zhang SM, Loker ES (2001) Structure of two FREP genes that combine IgSF and fibrinogen domains, with comments on diversity of the FREP gene family in the snail Biomphalaria glabrata. *Gene* 269:155–165.
60. Hanington PC, et al. (2010) Role for a somatically diversified lectin in resistance of an invertebrate to parasite infection. *Proc Natl Acad Sci USA* 107:21087–21092.
61. den Boon JA, Diaz A, Ahlquist P (2010) Cytoplasmic viral replication complexes. *Cell Host Microbe* 8:77–85.
62. Khatua AK, Taylor HE, Hildreth JE, Popik W (2009) Exosomes packaging APOBEC3G confer human immunodeficiency virus resistance to recipient cells. *J Virol* 83:512–521.
63. de Hoog GS, et al. (2011) Waterborne Exophiala species causing disease in cold-blooded animals. *Persoonia* 27:46–72.
64. Metzger MJ, et al. (2016) Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature* 534:705–709.
65. Nicolussi A, et al. (2018) Secreted heme peroxidase from Dictyostelium discoideum: Insights into catalysis, structure, and biological role. *J Biol Chem* 293:1330–1345.
66. Mungpakdee S, et al. (2014) Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. *Genome Biol Evol* 6:1408–1422.
67. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
68. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23:205–211.
69. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European molecular biology open software suite. *Trends Genet* 16:276–277.
70. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33:W244–W248.
71. Neuwald AF, Altschul SF (2016) Bayesian top-down protein sequence alignment with inferred position-specific gap penalties. *PLOS Comput Biol* 12:e1004936.
72. Cole 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:e9490.
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.