Supporting information appendix for

Diversification of AID-APOBEC-like deaminases in metazoa: identification of novel clades and widespread roles in immunity

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

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

[b]Department of Developmental Immunology, Max-Planck Institute of Immunobiology and Epigenetics, 79108 Freiburg, Germany.

*Author for correspondence: Aravind L (aravind@ncbi.nlm.nih.gov)

Contents:

Table S1. Phyletic distribution and Clade-specific synapomorphies of various novel AADs

SI text: SI Methods

Fig. S1. Multiple sequence alignment of AID/APOBEC clades shown in Fig.1A

Fig. S2. Phyletic distribution of AID/APOBEC-like proteins.

Fig. S3. Multiple Sequence alignment of AID

Fig. S4. Multiple sequence alignment of APOBEC1

Fig. S5. Multiple sequence alignment of APOBEC2

Fig. S6. Multiple sequence alignment of APOBEC3

Fig. S7. Multiple sequence alignment of NAD1

Fig. S8. Multiple sequence alignment of NAD2

Fig.S9. Multiple sequence alignment of Lophotrochozoan-Echinoderm-Cnidarian clade of AID/APOBECs closely related to the core vertebrate clade

Fig. S10. Multiple sequence alignment of Lophotrochozoan-Echinoderm clade of AID/APOBEC-like deaminases showing novel Zn chelation site involving loop-1 and loop-7

Fig. S11. Multiple sequence alignment of the Nematode clade of AID/APOBEC-like deaminases

Fig. S12. Multiple sequence alignment of the Arthropoda clade of AID/APOBEC-like deaminases.

Fig. S13. Multiple sequence alignment of APOBEC4

Fig. S14. Multiple sequence alignment of APOBEC4-Cnidarian/Algae clade showing their shared synapomorphies.

Fig. S15. Multiple sequence alignment of SNAD1

Fig. S16. Multiple sequence alignment of SNAD2

Fig. S17. Multiple sequence alignment of SNAD3

Fig. S18. Multiple sequence alignment of SNAD4

Fig. S19. Multiple sequence alignment of the Dictyosteliid clade of AID/APOBEC-like deaminases

Fig. S20. Multiple sequence alignment of the Bacterial clade of AID/APOBEC-like deaminases

Fig. S21. Supplementary version of the Fig. 2B showing lineage specific expansions (LSEs) of AADs

Fig. S22. Positional entropy bar graphs of all AAD clades.

Fig. S23. Phylogenetic trees and multiple sequence alignments of AADs showing intra-individual variations and pseudogenization

Fig. S24. A schematic representation of AADs showing mosaicism/potential recombination patterns and the corresponding multiple sequence alignments.

Fig. S25. Multiple sequence alignment showing LSEs of the DYW clade of deaminases in dinoflagellates

To access the raw data including protein sequences, phyletic distribution and multiple sequence alignments in text format refer ftp://ftp.ncbi.nlm.nih.gov/pub/aravind/AAD/apobec.html

Table S1. Phyletic distribution and Clade-specific synapomorphies of various novel AADs

Clades	Phyletic distribution	Synapomorphies
Secreted novel AID/APOBEC like deaminases 1 (SNAD1)	Ray-finned fishes, sporadic in amphibia, cryptodira, and lepidosauria, 2 copies in coelacanth, a one in *Ornithorhynchus* (platypus)	Signal peptide; intervening helix between Str-2 and 3; neomorphic Str-6 and Str-7 instead of canonical Hel-5 and 6 after Str-5; "C---CxxC---C" signature with the first C at the end of Str-5, CxxC in Str-6 and terminal C in Str-7; Cx_3C active site motif in loop next to Str-3; conserved cysteines in the loop next to Str-1 and at the beginning of Str-3 that might be involved in disulfide linkages; Several members harbor long substrate binding loop-7
Secreted novel AID/APOBEC like deaminases 2 (SNAD2)	Ray-finned fishes	Signal peptide; intervening helix between Str-2 and 3; novel Str-6 and 7 instead of canonical Hel-5 and 6 after Str-5; "C---CxxC---C" signature shared with SNAD1; Cx_nC active site motif in loop next to Str-3 (n=2-3); conserved tyrosine in middle of Str-3; highly variable loop-3 containing 4-64 residues and three cysteines that might stabilize the loop by disulfide linkages; C in Str-1 (only in minority, substituted by S)
Secreted novel AID/APOBEC like deaminases 3 (SNAD3)	Ray-finned fishes	Secondarily loss of signal peptide; extended bi-helical N-terminal segment before canonical Hel-1; novel (α+β) domain insert between the core Str-5 and Hel-5 containing four helices and two strands. This novel insert contains mostly conserved cysteines in Hel-1, Hel-2 and Hel-3 and Hel-4 that are predicted to form disulfide bridges; highly variable Cx_nC active site motif (n=2 to 15); conserved aromatic (Y/F) and a T in Str-3 before Cx_nC signature; highly variable loop-1
Secreted novel AID/APOBEC like deaminases 3 (SNAD4)	Porifera; found in *Amphimedon Queenslandica* (7) and *Haliclona tubifera* (18)	Signal peptide; Conserved Cx_2C active site signature; W in loop-7 instead of canonical Y; conserved W at the end of Str-5; stretch of negatively charged residues at the end of loop-7; Conserved charged residue (K/D(E)) before Cx_nC signature
Novel AID/APOBEC like deaminases 1 (NAD1)	Sporadic in Ray-finned and bony fishes, amphibia, lepidosauria, coelacanth, and metatherian mammals (marsupials)	Extended C-terminal segment after Hel-6, conserved FNN motif at the beginning of loop-1; conserved YNN motif in the beginning of loop-3. Conserved C in the middle of Str-1. Conserved Cx_2C active site signature; Conserved C in the middle of Hel-5
Novel AID/APOBEC like deaminases 2 (NAD2)	Sporadic in amphibia (*Microhyla fissipes*, *Rhacophorus dennysi*, *Rana catesbeiana*)	Conserved HY at the beginning of loop-1; HLTNT motif in loop-3; Conserved C in the middle of Str-1; Conserved Cx_2C active-site signature; Conserved C in the middle of Hel-5
Novel Lophotrochozoan Echinoderm and Cnidarian clade (Lop-Ech-Cni)	Found across several Lophotrochozoan Echinoderm and Cnidarian species; multiple paralogs in many species: *Patiria miniata* (12), *Ophiothrix spiculata* (8), *Evechinus chloroticus* (4), *Anthopleura elegantissima* (8) among others	Variable Cx_nC active site motif (n= 2 to 13 residues). Conserved NYS motif before the Cx_nC active site motif. Conserved WP before strand 1

AADs in Arthropods	Currently detected in 29 hexapod species and a crustacean *Proasellus parvulus*; Multiple losses across hexapods; found as single copy in 19 species, and multiple paralogs in 10 species: *Coptotermes gestroi* (6), *Periplaneta* americana (4), *Atelura formicaria* (4) among others	Partially conserved C at the beginning of Str-1, Canonical Y in loop-7 is often substituted by other aromatic residues F/H in several homologs; Conserved Cx_2C signature; Partially conserved C at the end of Str-5; Partially conserved C in the loop after Hel-3
AADs in Nematodes	Found across 35 species, single copies in most species	Nematode versions are highly conserved across species and are often fused to their N-terminus to eight repeats of a CxC-like domain. Conserved C in loop-1; Conserved Cx_4C active site signature; Extended loop-5 constitutes two additional conserved Cysteines; Conserved Cysteine at the beginning of Hel-3; Conserved Cysteine at the middle of Hel-4 and Hel-6; Extended C-terminal regions with additional helices after Hel-6. All nematode sequences are closely related
Novel Lophotrochozoan Echinodermata (Lop-Ech-Zn2 novel metal chelating clade: loop-1 and loop-7)	Found in several Lophotrochozoans and Echinoderms; expansions in many species including Strongylocentrotus purpuratus (15), *Crassostrea gigas* (13), *Cipangopaludina cathayensis* (9) among others	In addition to the active site Zn chelating site, this clade forms a novel Zn chelating site with residues involving loop-1 and loop-7. Conserved H in the beginning of loop-1 or end of Hel-1 (GLH motif), Conserved SC in loop-7, followed by two additional positions in loop-7 that constitute either C or H forms a neomorphic H-C-C/H-C/H Zn chelating signature. Conserved WP motif in loop1 before Str-1. Lengthy loop-7 contributing to novel Zn chelation can vary between 9 to 28 residues. Conserved DW in helix-5. Cx_nC active site motif (n=2 to 6 residues)
APOBEC4 and Cnidarian algae clade (novel metal chelating clade: loop-1/Helix-1 and loop-3)	APOBEC4 is found across most vertebrates including lamprey. Cnidarian members of the clade are found across several species including LSEs in some. Members of this clade are also found in haptophytes *Prymnesium parvum* (2 copies) and *Emiliania huxleyi* (5 copies) as well as a green algae *Micromonas commoda*	Conserved C in loop-1/Hel-1, Conserved H in loop-1/Hel-1, Conserved NC motif at the beginning of loop-3; RxxxD motif in Hel-6. In addition to the active site Zn chelating site, this clade forms a neomorphic Zn-chelating site with residues from loop-1/Hel-1 and loop-3. They do show some differences in the configuration. Thus, APOBEC4's neomorphic Zn chelating site is formed by three residues from loop-1 (C-C-H) and another C from loop-3 (NC motif) next to the characteristic loop-3 asparagine. The equivalent neomorphic Zinc-chelating residues from Cnidarian-algae homologs is formed by C and H from end of helix-1, C from loop-1 and another C from loop-3 (NC motif). The Cnidarian-algae homologs also encodes an additional predicted Zn-chelating site formed between N-terminal stretch containing CxxC-----H and another C from loop-3. Variable length of Cx_nC active site signature (n=2 to 25 residues)
Secreted Dictyosteliid AADs	LSEs in Dictyosteliida: *Dictyostelium fasciculatum* (7 copies), *Tieghemostelium lacteum* (6 copies). *Acytostelium subglobosum* LB1 has a lone member	Signal peptide, Distinct C-terminal domain that are likely to form disulfide bridges. An insert after Helix-1 and strand-1. Conserved W in loop-7, equivalent to the signature Y. Conserved polar/small residue (S/Q) in loop-1. Conserved polar (Q) at the beginning of loop-3

SI Methods

S. purpuratus, *C. gigas* and *L. antina* gDNA extraction and WGS

Live *C. gigas* samples were obtained from Christian Buschbaum (Alfred-Wegner Institute, Sylt, Germany). Shells were opened and body sliced into smaller pieces with a fresh scalpel before being snap frozen in liquid nitrogen and ground using a mortar and pestle. Half of the material was put in 10ml of TriReagent (Sigma) for RNA extraction (see below) and the remaining half used for gDNA extraction. For gDNA extraction the material was re-suspended in 30ml "shellfish extraction buffer" (10mM Tris pH8, 100mM EDTA, 200mM NaCl and 0.5% SDS) and 400µl 20mg/ml proteinase K was added. The sample was digested at 55°C overnight. Whole *L. anatina* samples were provided by Elena Temereva (Moscow State University, Moscow, Russia)., stored in RNALater® (Sigma). The sample was briefly washed in PBS before being processed as for *C. gigas*. *S. purpuratus* coelemocytes stored in RNALater® were provided by Eric Davidson (California Institute of Technology, Pasadena, USA). The sample was diluted with PBS and cells were pelleted at 3000xg for 5 minutes at room temperature and re-suspended in 600µl "sea urchin sperm isolation buffer" (20mM NaCl, 50mM Tris pH8, 20mM EDTA). 3µl of 20mg/ml proteinase K and 15µl 20% SDS was added and mixed by inversion and incubated for 1 hour at 65°C and followed by overnight at 37°C. gDNA of all proteinase K digests was extracted using phenol-choloform, including an RNAase treatment step before re-precipitation and elution in TE buffer. Concentration was measured on a Qbit 3.0 (Invitrogen) and genomic integrity was assessed on an agarose gel. Both *C. gigas* and *S. purpuratus* showed fully intact gDNA whereas *L. anatina* showed some degradation. Sequencing libraries were made from a minimum of 1µg of total gDNA using the Illumina TruSeq® DNA PCR-Free LT Sample Preparation Kit (FC-121-3001). Libraries were sequenced at 2x250bp paired-end to a depth of 150 million reads using the HiSeq2500 Rapid Mode (Illumina).

Generation of BLAST databases from WGS.

Illumina adapter sequences were trimmed off the read ends using cutadapt in paired-end mode. Low quality bases were removed using prinseq with minimum base quality 20 in a window of 3 bases with a step size of 2 bases, retaining reads of at least 30nt in length. For lower quality libraries, window and step size were increased. Reads with mean quality below 26 were removed as appropriate. Forward and reverse mates were merged using flash, allowing for up to 300 bases overlap. Spaces in read names were replaced with underscores using sed. The resulting merged as well as unmerged read files were formatted into nucleotide blast databases allowing for read name parsing.

***S. purpuratus, C. gigas* and *L. antina* RNA extraction and RNA-seq**

Samples were processed as described above and total RNA was extracted using TriReagent (Sigma) and treated with RNAase-free DNAse (Roche) before being re-extracted in the same way. The RNA-seq library was made from 300ng – 3000ng total RNA using the Illumina TruSeq® stranded mRNA library prep kit (RS-122-2101) utilizing polyA selection to enrich for mRNA. Libraries were then amplified for 5 cycles. Libraries were sequenced at 2x100bp paired-end to a depth of 100 million reads.

Generation of *de* novo transcriptomes

Adapter sequences were trimmed with Trim Galore! (Galaxy Version: 0.4.2) with subsequent read-ends below a quality score of 15 also removed and resulting reads shorter than 20 nucleotides also discarded. Reads were directly assembled *de novo* using Trinity (1) (Galaxy Version: 2.4.0.0) using default parameters for stranded libraries using paired end reads. We included *in silico* fragment normalisation to 30x coverage and set a minimum contig length cut-off of 200 nucleotides. The best ORF per transcript was extracted, setting a minimum cut-off of 30 amino acids and identical ORFs at the amino acid level were collapsed using CD-HIT PROTEIN (2) (Galaxy Version: 1.3) with default parameters and an identity threshold of 0.99. A second, more stringent set of transcripts was also identified where "Good" transcripts from the *de novo* assembly were identified using TransRate (3) (Linux version 1.0.3) using default parameters. ORFs were extracted and collapsed as before.

Organism	Trinity Transcripts	Initial Output		Transrate Output		
		ORF	Unique ORF	Transrate Good	ORF	Unique ORF
C. gigas	350337	158676	116046	191968	27421	25638
L. antina	94726	17665	14675	62620	10177	9858
S. purpuratus	124787	22911	17187	81902	12224	11667

1. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7):644-652.
2. Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658-1659.
3. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, & Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research* 26(8):1134-1144.

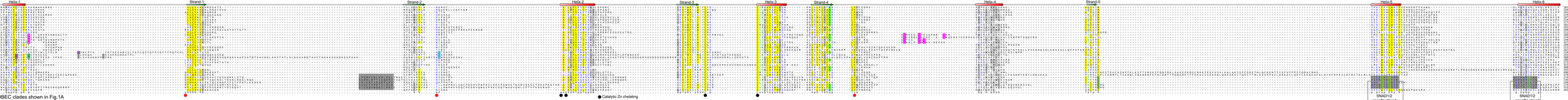

Fig. S1. Multiple sequence alignment of AID/APOBEC clades shown in Fig.1A

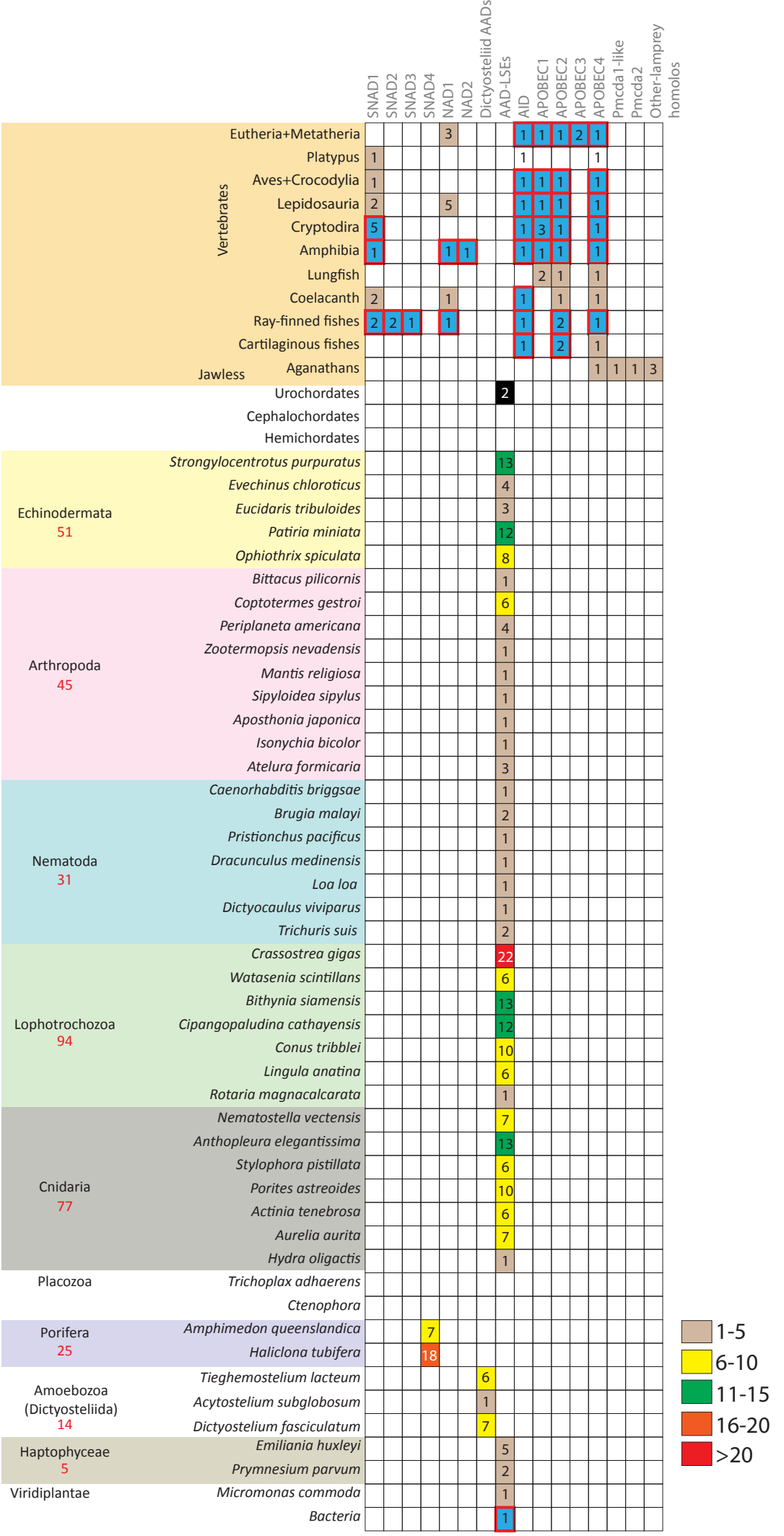

Fig. S2. Phyletic distribution of AID/APOBEC-like proteins. Blue boxes with red outline represent the median values within the lineage, colored boxes with no red outline show the absolute protein counts. Absolute counts range between 1-5, 6-10,11-15, 15-20, and >20 are colored pale brown, yellow, green, orange and red respectively. Black box for urochordates represent inactive versions. The oveall count shown in red below the names of the major lineages (on the left) are an estimate at 95% sequence identity. In mammals, NAD1 is present only in Metatherians (Marsupials), whereas APOBEC3 is present only in Eutherians. To access the all protein sequences identified in this study refer ftp://ftp.ncbi.nlm.nih.gov/pub/aravind/AAD/apobec.html

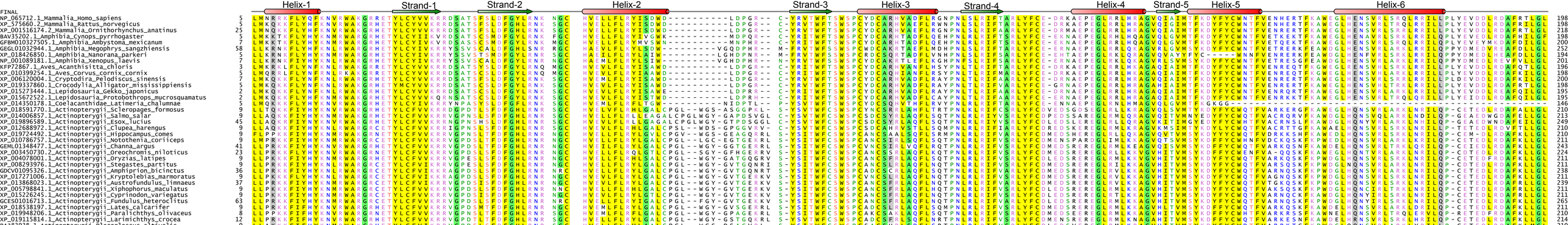

Fig. S3. Multiple sequence alignment of AID

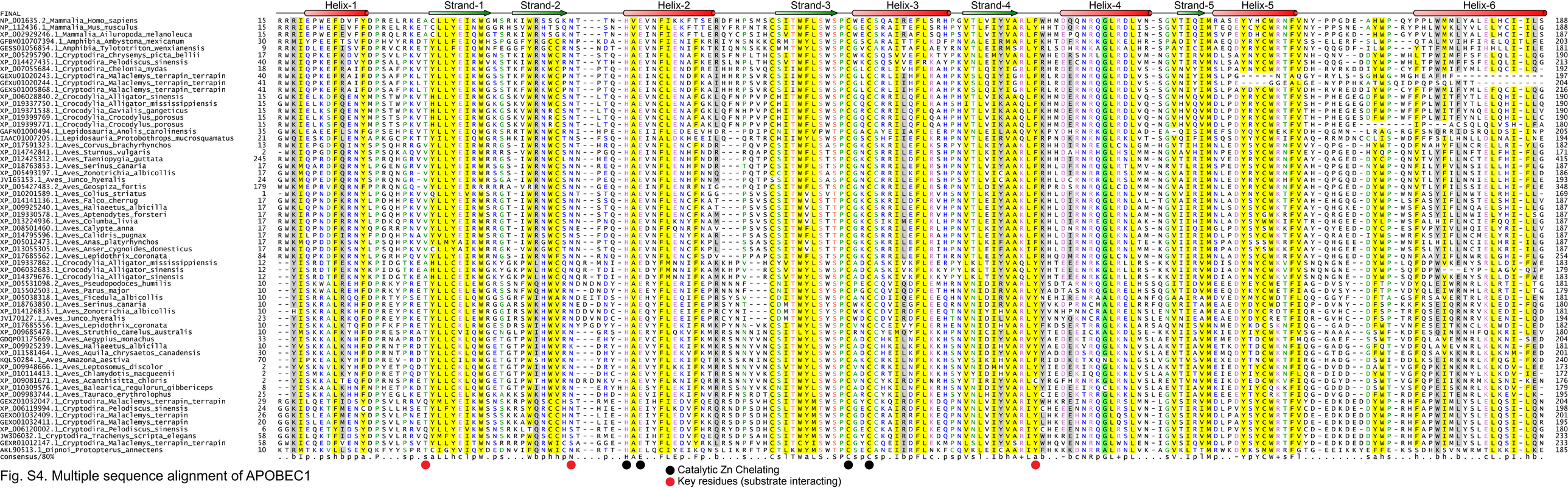

Fig. S4. Multiple sequence alignment of APOBEC1

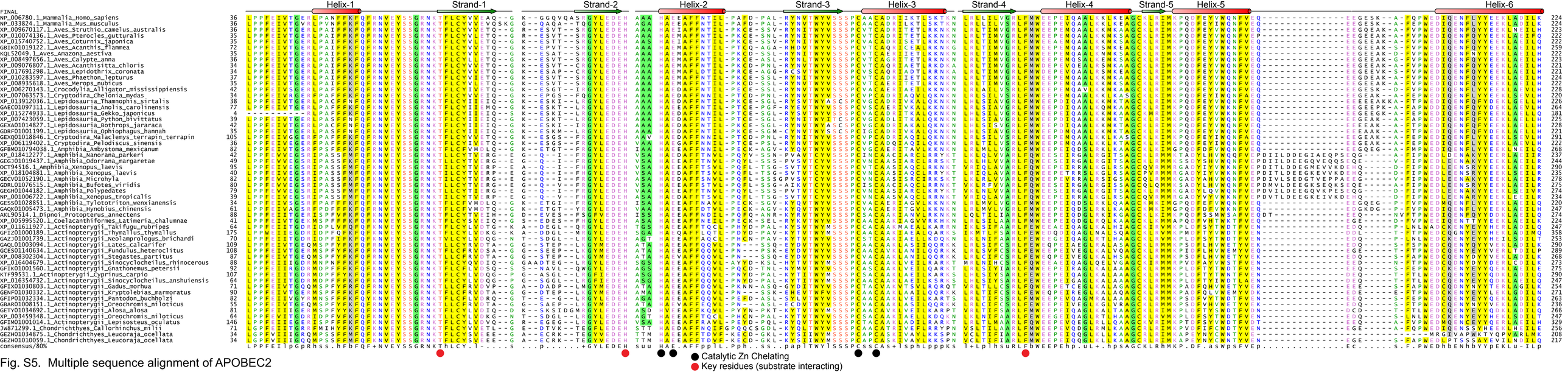

Fig. S5. Multiple sequence alignment of APOBEC2

Fig. S6. Multiple sequence alignment of APOBEC3

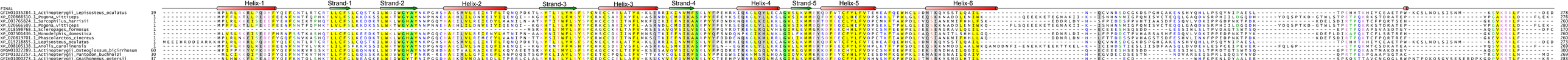

Fig. S7. Multiple sequence alignment of NAD1

Fig. S8. Multiple sequence alignment of NAD2

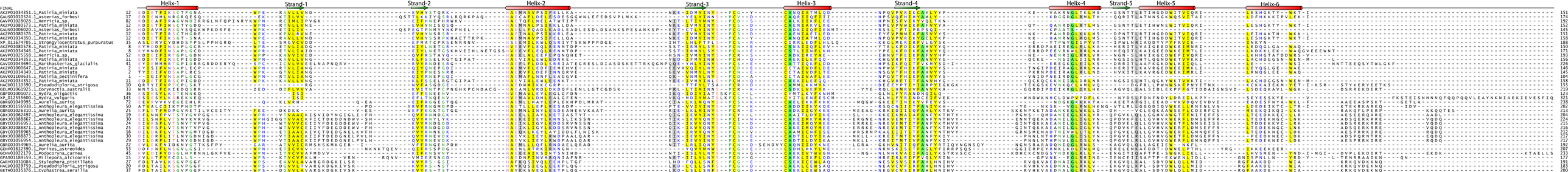

Fig. S9. Multiple sequence alignment of Lophotrochozoan-Echinoderm-Cnidarian clade of AID/APOBECs closely related to the core vertebrate clade

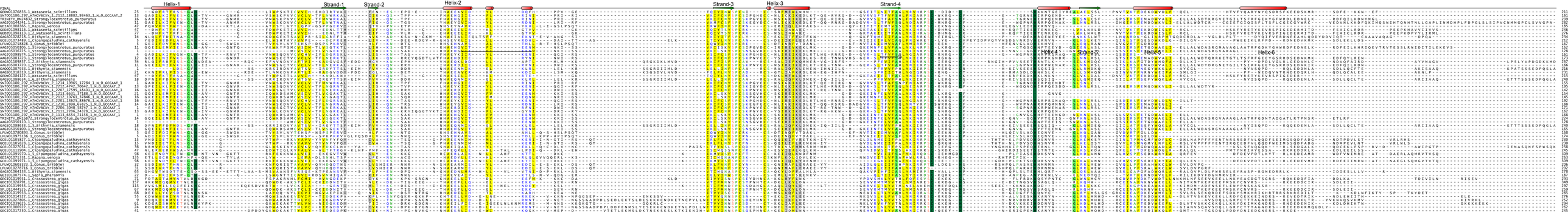

Fig. S10. Multiple sequence alignment of Lophotrochozoan-Echinoderm clade of AID/APOBEC-like deaminases showing novel Zn chelation site involving loop-1 and loop-7

● Novel Zn chelating residues

● Catalytic Zn chelating

Fig. S11. Multiple sequence alignment of the Nematode clade of AID/APOBEC -like deaminases

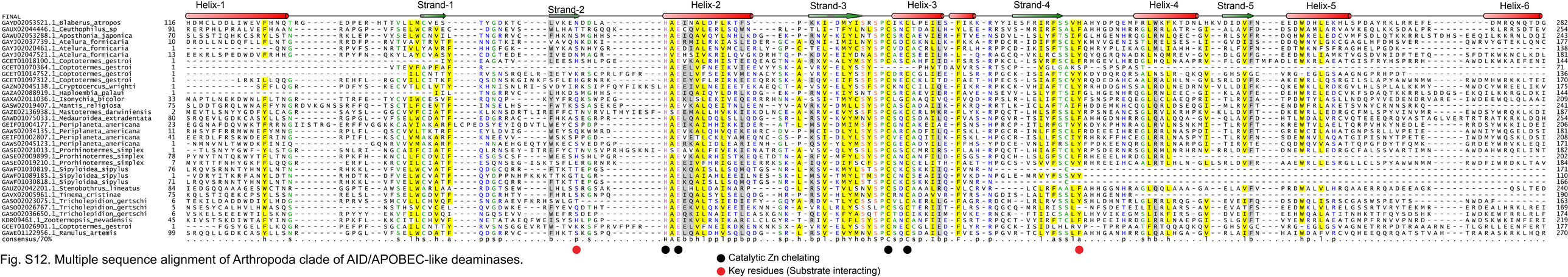

Fig. S12. Multiple sequence alignment of Arthropoda clade of AID/APOBEC-like deaminases.

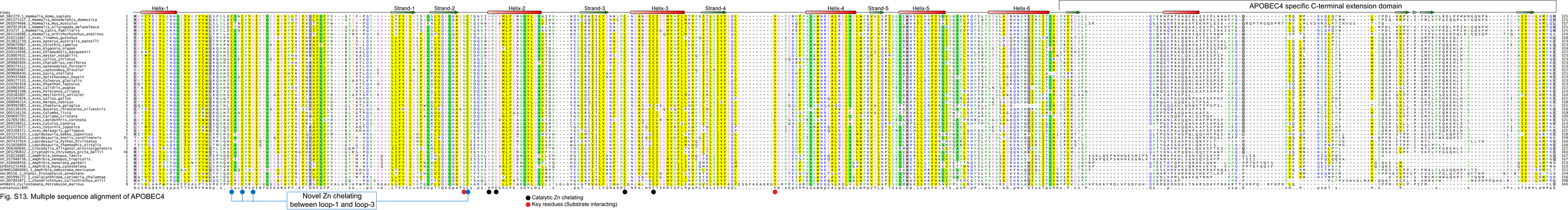

Fig. S13. Multiple sequence alignment of APOBEC4

Fig. S14. Multiple sequence alignment of APOBEC4-Cnidarian/Algae clade showing their shared synapomorphies.

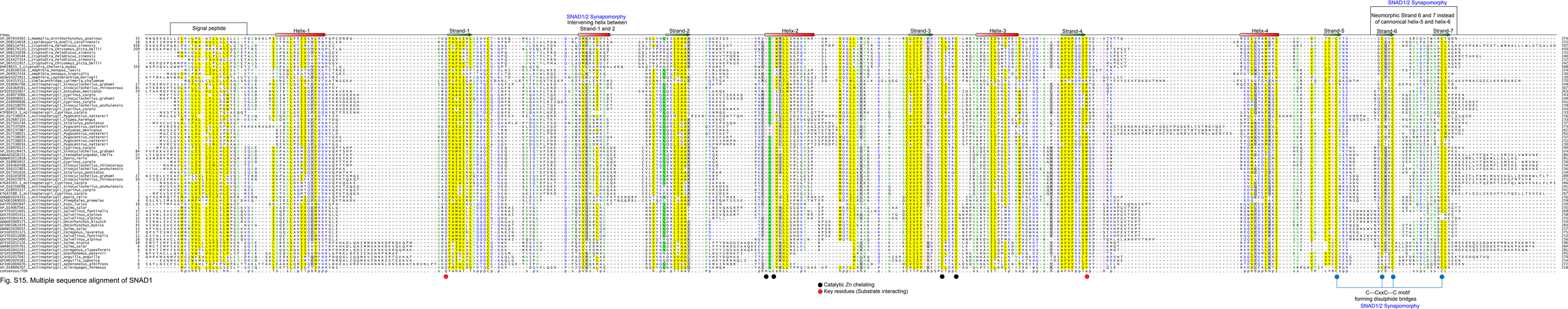

Fig. S15. Multiple sequence alignment of SNAD1

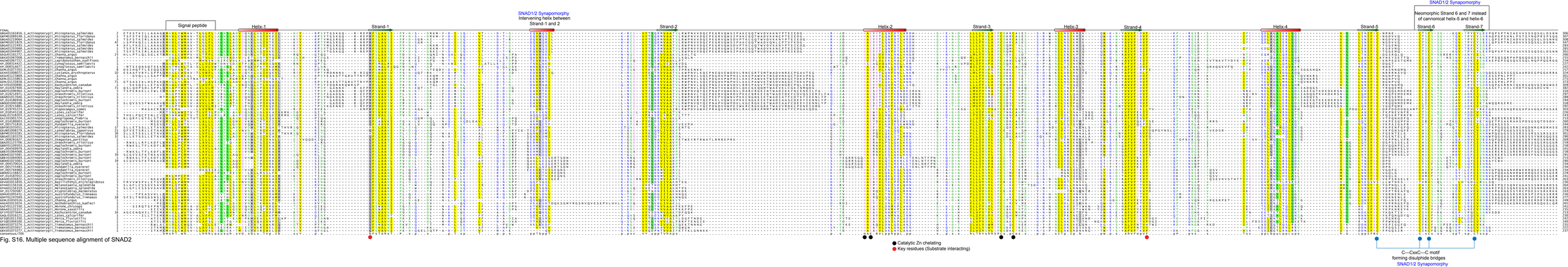

Fig. S16. Multiple sequence alignment of SNAD2

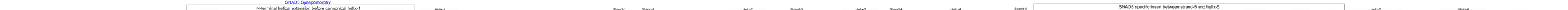

Fig S17. Multiple sequence alignment of SNAD3

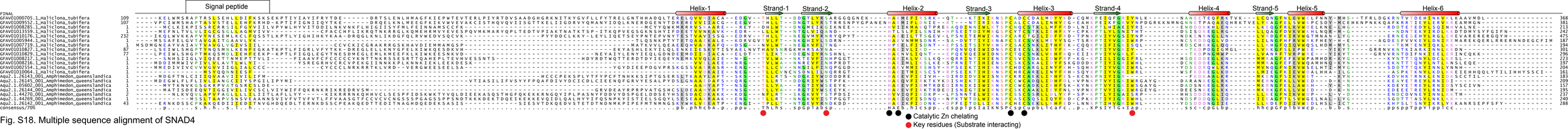

Fig. S18. Multiple sequence alignment of SNAD4

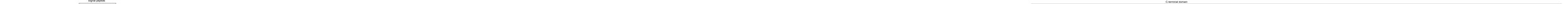

Figure S19. Multiple sequence alignment of secreted Dictyostelida AADs.

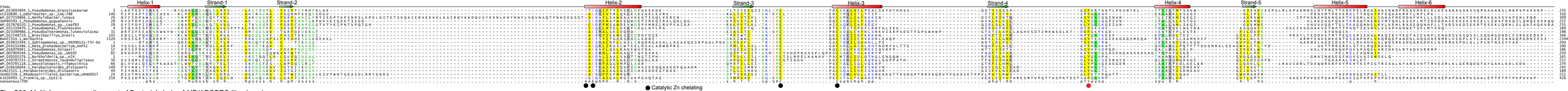

Fig. S20. Multiple sequence alignment of Bacterial clade of AID/APOBEC-like deaminases

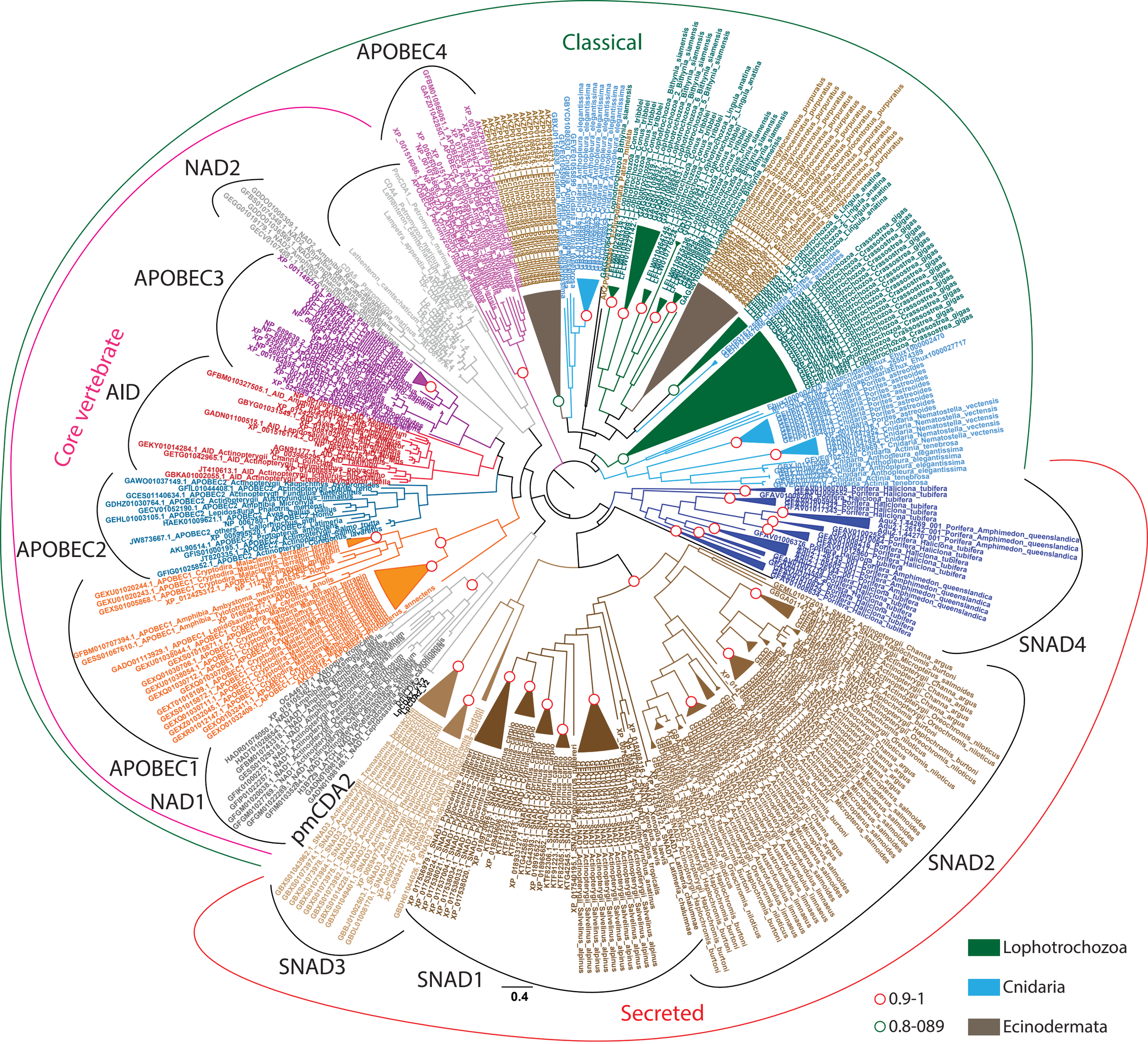

Fig. S21. Supplementary version of the Fig. 2B showing lineage specific expansions (LSEs) of AADs

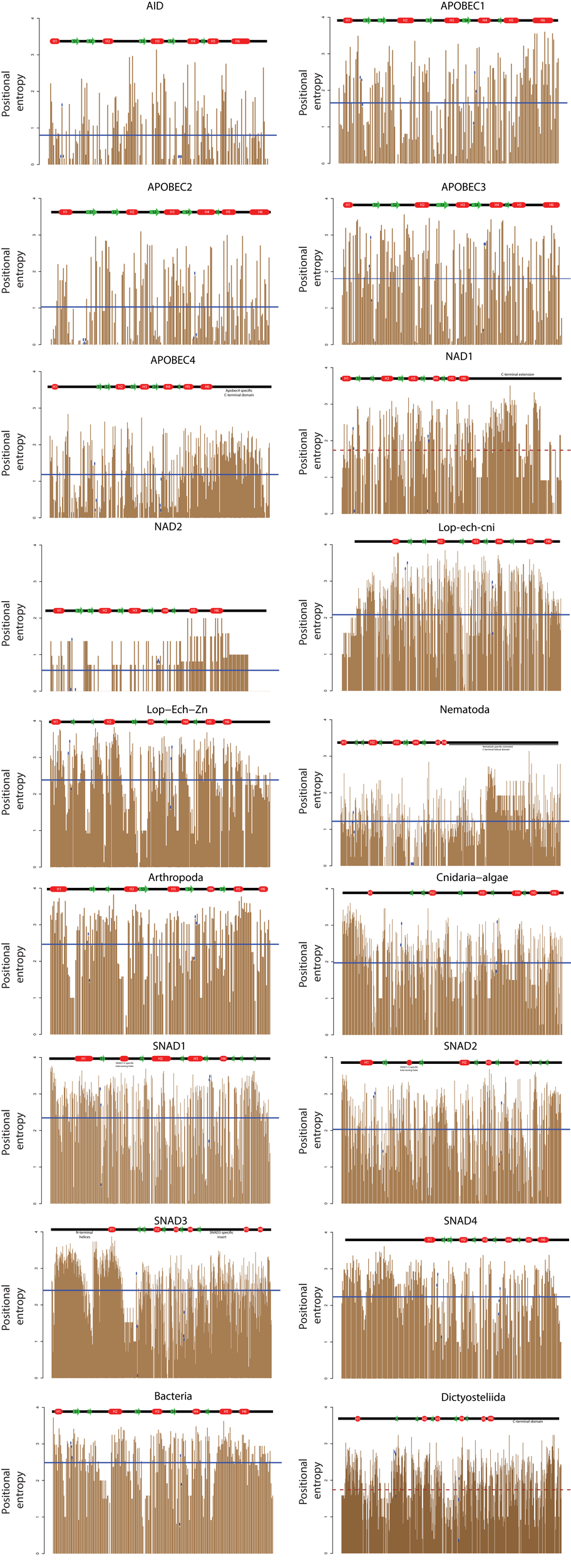

Fig. S22. Positional entropy values plotted for various AAD clades with mean entropy (blue horizontal line) with secondary structure on the top. Loop-1 and 7 residues shown in Fig. 3E are marked with blue dots.

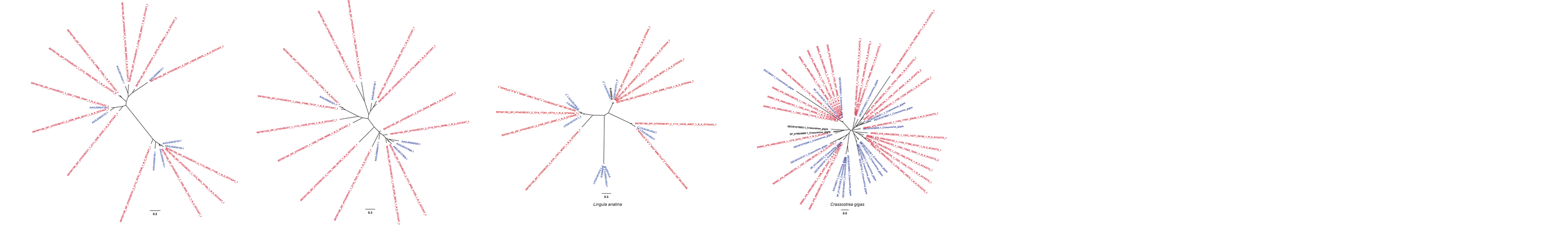

Fig. S23a. Tree topologies showing intra-individual variation of paralogous AADs

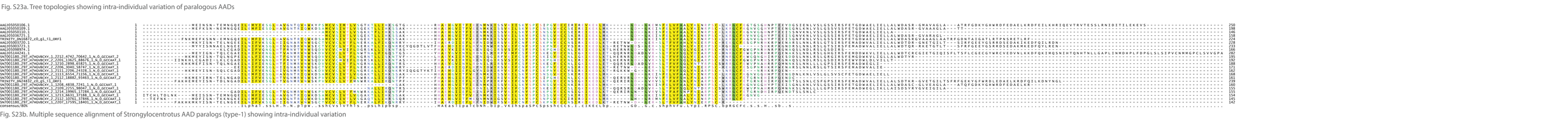

Fig. S23b. Multiple sequence alignment of Strongylocentrotus AAD paralogs (type-1) showing intra-individual variation

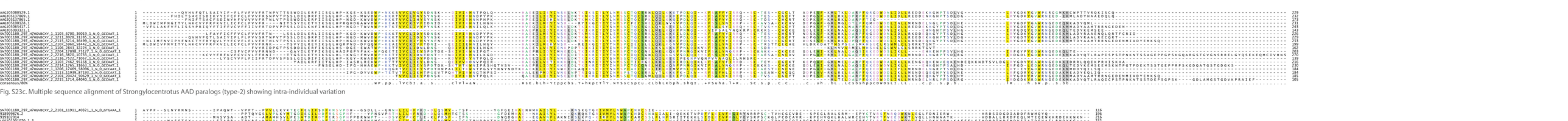

Fig. S23c. Multiple sequence alignment of Strongylocentrotus AAD paralogs (type-2) showing intra-individual variation

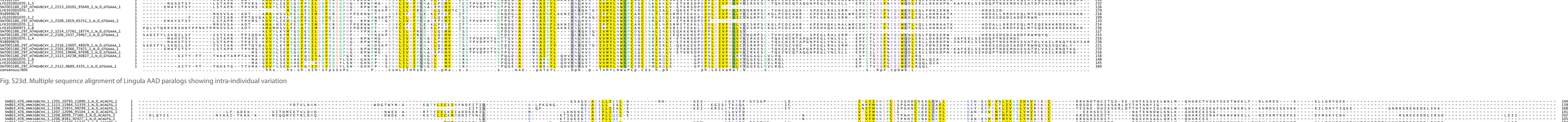

Fig. S23d. Multiple sequence alignment of Lingula AAD paralogs showing intra-individual variation

Fig. S23e. Multiple sequence alignment of Crassostrea AAD paralogs showing intra-individual variation

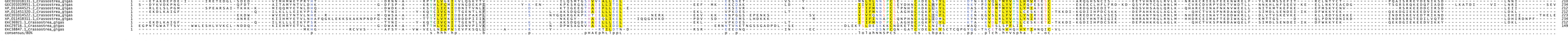

Fig. S23e. Multiple sequence alignment of Nematostella AADs showing pseudogenization. * denotes stop codon

Fig. S23. Phylogenetic trees and multiple sequence alignments of AADs showing intra-individual variations and pseudogenization

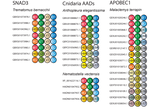

Fig. S24a. Schematic representation of AADs showing mosaicism or potential recombination pattern and C-terminal variation. Abbreviations: nt: N-terminal segment ; ct C-terminal segment ; c: Core segment ; s: Shared segment.

Fig. S24b. Multiple sequence alignment of *Trematomus* SNAD3 showing mosaicism.

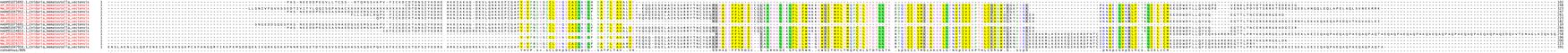

Fig. S24c. Multiple sequence alignment of *Nematostella* AADS showing mosaicism.

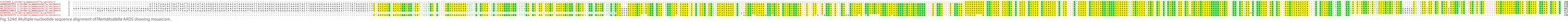

Fig. S24d. Multiple nucleotide sequence alignment of *Nematostella* AADS showing mosaicism.

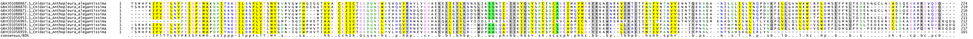

Fig. S24e. Multiple sequence alignment of *Anthopleura* AADS showing mosaicism.

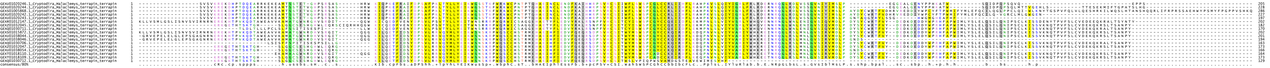

Fig. S24f. Multiple sequence alignment of APOBEC1 paralogs from turtle *Malaclemys* showing C-terminal variation

Fig. S24. A schematic representation of AADs showing mosaicism/potential recombination patterns and the corresponding multiple sequence alignments

Fig. S25. Multiple sequence alignment showing LSEs of the DYW clade of deaminases in dinoflagellates