

# Supplementary Information for

A family of unusual immunoglobulin superfamily genes in an invertebrate histocompatibility complex

Aidan L. Huene<sup>1,2</sup>, Steven M. Sanders<sup>1,2</sup>, Zhiwei Ma<sup>1,2</sup>, Anh-Dao Nguyen<sup>3</sup>, Sergey Koren<sup>3</sup>, Manuel H. Michaca<sup>1,2</sup>, Jim C. Mullikin<sup>3,4</sup>, Adam M. Phillippy<sup>3</sup>, Christine E. Schnitzler<sup>5,6</sup>, Andreas D. Baxevanis<sup>3</sup>, and Matthew L. Nicotra<sup>1,2,7</sup>

Affiliations:

<sup>1</sup> Starzl Transplantation Institute, University of Pittsburgh, Pittsburgh, PA, USA

<sup>2</sup> Center for Evolutionary Biology and Medicine, University of Pittsburgh, Pittsburgh, PA, USA

<sup>3</sup> Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

<sup>4</sup> NIH Intramural Sequencing Center, National Institutes of Health, Rockville, MD, USA
 <sup>5</sup> Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL, USA

<sup>6</sup> Department of Biology, University of Florida, Gainesville, FL USA

<sup>7</sup> Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA

\*Corresponding Author: Matthew L. Nicotra **Email:** <u>matthew.nicotra@pitt.edu</u>

# This PDF file includes:

Supplementary Materials and Methods Figures S1 to S23 Tables S1 to S9

Legends for Datasets S1 to S14

SI References

# Other supplementary materials for this manuscript include the following:

Datasets S1 to S14

### **Supplementary Information Text**

#### Material and Methods

#### Sequencing and assembly of the genome of an ARC homozygous animal

Colony 236-21 was maintained on glass microscope slides in 38 liter aguaria filled with artificial seawater (Reef Crystals) as previously described (1). It was starved three days prior to nucleic acid extraction. Tissue was scraped from the slide with a sterile razor blade and snap-frozen by transferring it to a mortar filled with liquid nitrogen. The frozen tissue was then ground into a fine powder with a pestle. UEB1 buffer (7 M urea, 0.3125 M NaCl, 0.05 M Tris-HCl, 0.02 M EDTA, and 1% w:v N-lauroylsarcosine sodium salt) was added to the mortar, where it froze. The frozen UEB1-tissue mixture was ground into a powder and transferred to a 50 ml centrifuge tube containing room temperature UEB1 buffer. This was mixed by gentle inversion. An equal volume of equilibrated phenol:chloroform:isoamyl alcohol (25:24:1) was added and mixed by gentle inversion. This was centrifuged for 10 minutes at 3000 x g. The aqueous layer was transferred to a 15 ml centrifuge tube with a wide bore pipette tip. Total nucleic acid was precipitated by adding 0.7 volume isopropyl alcohol. Precipitated nucleic acid was then spooled onto a pipette tip and transferred to a clean 15 ml tube, where it was washed twice with 70% EtOH and twice with 100% EtOH. The precipitated material was then gently brought to the bottom of the tube by briefly centrifuging, air dried, and immediately resuspended in 1X TE (10 mM Tris-HCl, pH 8.0; 1 mM EDTA, pH 8.0). RNA was then digested by adding RNAses (RNAse cocktail, Ambion, #AM2286) and incubating at 37°C for 15 minutes. DNA was then extracted by adding 1 volume equilibrated phenol:chloroform:isoamyl alcohol, centrifugation at 12,000 x g, and transfer of the aqueous layer to a new tube. This was followed by precipitation with 2.5 volumes of 100% ethanol and 1/10 volume 5 M sodium acetate (pH 5.2). The precipitate was pelleted, washed with 70% ethanol, and resuspended with 1X TE. The resuspended DNA was then stored at -20°C.

PacBio and Illumina libraries were constructed and sequencing performed at the NIH Intramural Sequencing Center (NISC) via a whole-genome shotgun approach. Both the high-throughput Illumina HiSeq2500, run as 250 base paired-end reads, and PacBio RSII long-read sequencing platforms were used. Filtered subreads from two PacBio libraries (a total of 37 SMRT cells) were corrected with the Celera Assembler version 8.3r2 (2). Specifically, the PBcR.pl script was used with parameter sensitive=1, which is recommended for datasets with <50x coverage, increases MHAP sensitivity, and uses the slower but more accurate pdbagcon consensus algorithm to generate the corrected reads. The corrected reads were assembled with runCA.pl parameters batOptions="-el 5000 -eg 0.025 -Eg 4.00 -em 0.025 -Em 4.00 -o asm -RS -NS -CS -repeatdetect 6 150 15" which reduces the default error rate, increases the minimum overlap size, and increases the splitting thresholds. The resulting assembly was polished with the PacBio reads using the ArrowGrid parallel wrapper (3) followed by polishing with the Illumina short read data using the PilonGrid parallel wrapper (4).

#### Assembly of the ARC

To assemble the full reference sequence for the ARC, NUCmer from the MUMmer package (v3.23) was used align the BAC contigs with the newly assembled whole genome sequence to identify the contigs which matched the known ARC sequence (5). First, the query and reference sequences were aligned using NUCmer (nucmer -p <output.file> <reference.file> <query.file>). The resulting file was then filtered (delta-filter) to only show matching hits in one direction on the strands (-r) and to remove all hits less than 1000 base pairs (-I #). Finally, the output was appended into a tab-delimited file (-T) sorted by the reference sequence (-r), with a minimum length of 1 kb or 10 kb (-L #), the sequence length (-I), and the percent coverage between two sequences (-c). The tabular files were manually inspected to assess overlapping contigs. Overlapping regions were then inspected by alignment with BLAST+ version 2.6.0 (6) and dot plots generated in YASS (7). The genome assembly and BAC sequences were then merged to create a reference sequence of the ARC-F haplotype (File S1).

#### RNA extraction, sequencing, and mapping

Thirty polyps were severed from colony 236-21 with a scalpel, moved to an eppendorf tube and briefly centrifuged. Remaining water was removed with a pipette. Tissue was immediately lysed with 0.5 mL of TRIzol (Invitrogen) and ground vigorously with a small pestle. Lysate was incubated for less than five minutes at room temp. One hundred  $\mu$ l chloroform was added and the tube was shaken vigorously for 15 seconds, followed by a three minute incubation at room temp. The sample was then centrifuged at 12,000 x g for 15 minutes at 4°C. RNA was then extracted from the aqueous phase with a PureLink RNA Mini Kit (Invitrogen). RNA quality and quantitation was assessed by Tapestation and Qubit, respectively, at the University of Pittsburgh Genomics Core. Final sample was frozen and stored at -80°C until sequencing by NISC. RNA-Seq libraries were constructed from 1  $\mu$ g RNA using the Illumina TruSeq Stranded mRNA kit. The resulting cDNA was fragmented using a Covaris E210 focused ultrasonicator. Library amplification was performed using 10 cycles to minimize the risk of over-amplification. The library was sequenced on an Illumina HiSeq4000 to generate 75 base paired-end reads.

To calculate expression levels of our annotated Alr genes, paired-end RNA-seq reads were mapped to the entire genome assembly using HISAT2 (8). The mapping was performed under the most stringent conditions (only concordant mappings with zero mismatches were kept) while allowing for multiple alignments. The resulting mapping file was processed and sorted using samtools (9) before proceeding to quantitation. Using the reference annotations of the Alr genes, transcript abundance of the Alr genes was estimated with Cufflinks (10). Abundance estimates were corrected for multiple read mappings.

### Annotation of *Alr* genes

*Alr* genes were annotated using Apollo (11) installed on a local computer running Ubuntu 18 LTS. Tracks displaying the results of BLASTX searches and RNAseq mapping were imported and used as a guide for manual annotation of *Alr* gene models. To generate BLAST results, repeats in the genomic sequences were first masked using the protein-based repeat masking option on the Repeatmasker website (<u>https://www.repeatmasker.org</u>) (12). Masked DNA sequences were then divided into 32 kb segments with 2 kb overlaps. These segments were used as BLASTX queries against a database of Alr1 and Alr2 proteins (to identify *Alr*-like sequences), and the swissprot database (to identify highly conserved genes). BLAST results were then concatenated, and a custom perl script was used to adjust their coordinates to those of the unsegmented genome sequence. To generate RNAseq alignments, the assembled RNA-seq dataset was aligned to the genome using HISAT2 (v2.1.0) through the Galaxy platform (8, 13). The parameters used RNAseq alignments included paired-end reads and no alignments for individual mates, and only 1 primary alignment. The output file (.bam) was then imported to Apollo for visualization during annotation.

#### Alr sequence comparisons

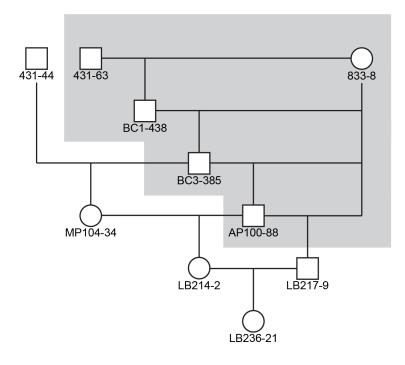
Alignments between AIr proteins were performed using MAFFT (14). The L-INS-i alignment strategy was used for all alignments except those involving only domains 1, 2, and 3, which used G-INS-i. Pairwise sequence alignments were done using the modified Needleman-Wunsch algorithm available in Jalview (15). Clustering was performed with CD-HIT (16) using the psi-cd-hit.pl script, with 20% sequence identity cutoff and 0.1 e-value cutoff. Neighbor joining trees were constructed in Jalview using the BLOSUM62 scoring matrix. Trees were visualized in iTOL (17), exported as scaled vector graphics files, and annotated in Adobe Illustrator.

### Protein sequence analysis

Signal peptides were predicted with SignalP 5.0 (18). Transmembrane helices were predicted with TMHMM 2.0 (19). Conserved protein domains were identified with the Pfam database (20) using HMMER3 (<u>http://hmmer.org/</u>). For domain prediction by HHpred, sequences were submitted to the MPI Bioinformatics Toolkit (21). The query MSA was generated via three iterations of HHblits against the Uniref30 database, with an e-value threshold of 1 x  $10^{-3}$  for inclusion. HHpred was then used to search the SCOPe70\_2.07 database.

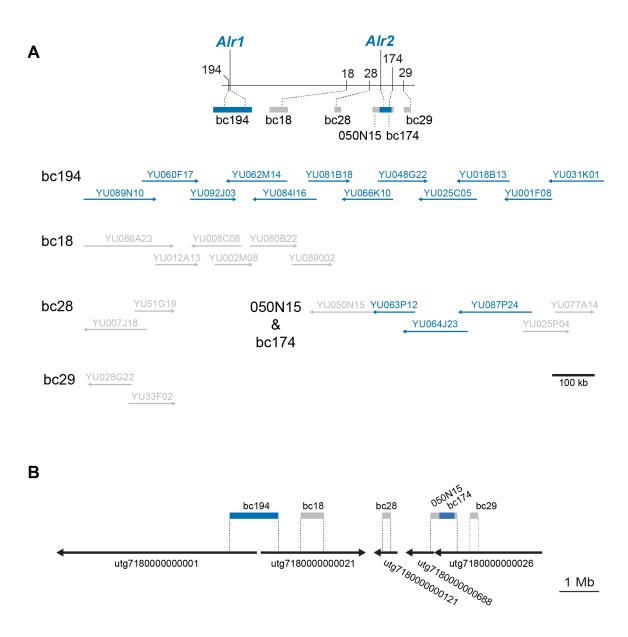
### Structural prediction and alignment

For single domain predictions, we generated a custom multiple sequence alignment which was submitted, along with the query sequence, to Colabfold via the "AlphaFold2\_mmseqs2" notebook, version 1.1 (22). The input multiple sequence alignment was generated as follows. The sequence of the query domain was aligned to the same domain type from all *bona fide* Alr proteins using MAFFT with the G-INS-i setting. This alignment was then submitted as a query to HHblits via the MPI Bioinformatics Toolkit, which was run for two iterations against the Uniref30 database, with an e-value threshold of 1 x 10<sup>-3</sup> for inclusion. A reduced representation alignment of the resulting Query MSA was then downloaded and submitted as a custom multiple sequence alignment to Colabfold. The secondary structure of each model was determined with STRIDE (23). The Alr structural model with the highest average pLDDT was then submitted to DALI (24) and PDBeFOLD (<u>https://www.ebi.ac.uk/msd-srv/ssm/</u>) (25) to identify similar structures in the PDB. For multi-domain predictions, the query sequence was submitted directly to Colabfold, with msa\_mode = "Mmseqs2 (Uniref + Environmental)". Models were visualized in Pymol 2.3 (26).



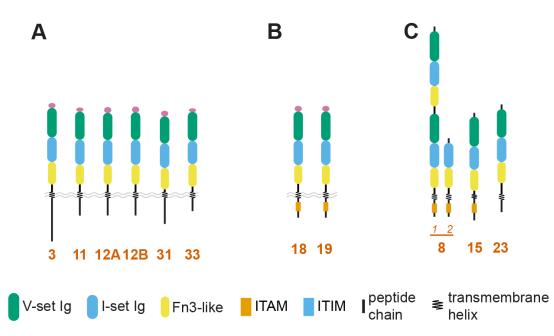
# Fig. S1. Pedigree of colonies used to generate ARC-F reference sequence.

The pedigree of colony LB236-21 can be recreated by concatenating previously published pedigrees (shaded area) (1, 27). Colony AP100-88 is from the mapping population in Powell *et al.* (27). Colony 431-44 is from the mapping population in Cadavid *et al.* (1).



# Fig. S2. Detail of ARC reference assembly.

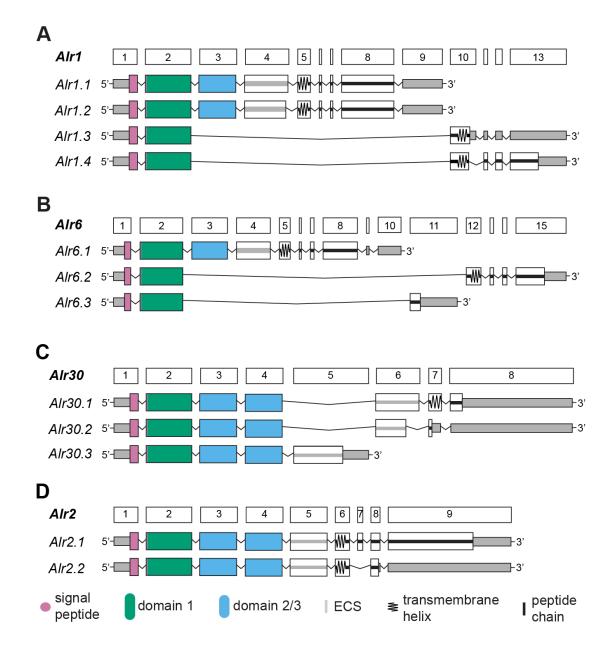
- (A) Minimum tiling path of sequenced BAC clones resulting from chromosome walks from five markers in the ARC linkage map. Clone names are indicated above an arrow indicating their orientation. Sequences reported in (28) or (29) are in navy blue. Unpublished sequences are in gray.
- (B) Overlap between BAC contigs and genome contigs.



### Fig. S3. Predicted domain architectures of putative Alr proteins.

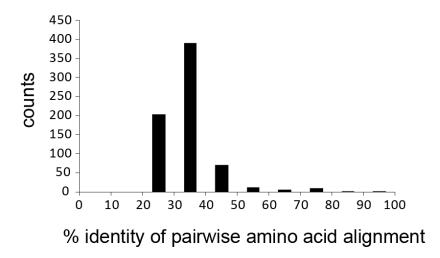
- (A) Domain architecture of proteins encoded by unexpressed putative genes.
- (B) Domain architecture of proteins encoded by partially expressed putative genes.
- (C) Domain architecture of proteins encoded by putative genes that are predicted to lack a signal peptide.

Final domain predictions and the presence of ITAM and ITIM motifs are indicated here and described later in the main text. In (A) and (B), signal peptides are expected to be cleaved, but are shown to indicate their presence.



# Fig. S4. Alternative splicing of Alr genes

In all panels, exons are colored according to the type of domain/region they encode. (A) alternative splicing of *Alr1*. (B) Alternative splicing of *Alr6*. (C) Alternative splicing of *Alr30*. (D) Alternative splicing of *Alr2*.



#### Fig. S5. Pairwise amino acid identities between Alr genes

Histogram of amino acid percent identities for pairwise alignments of *Alr* genes and putative genes. Alignments were performed using the modified Needleman-Wunsch algorithm available in Jalview.

Α		group 1	group 2	group 3 👩
	0		9 11 124 123	
В	3 	- 4 Cluster A	16 Cluster B 28 15 27 29 1000 1 1	Image: Cluster C         34         35           gap         gap         33         36           Image: Cluster C         34         35           Image: Cluster C         34         35           Image: Cluster C         33         36           Image: Cluster C         1         Mb
D	Alr16 Alr18 Alr19 Alr21	390         KHKGFIIFIKNPKKR-           383         KHKGFIIFIKNPKNQ-           378         KHKGYITVSINRKR           378         RKGYLVFKKSSNTT-           377         KRKVLKEIFKEKNH-	- LOHTKRRNDDNYEVTQI - SKTDVKDNIDDYEVAGT - PQSNENGS-DDYEETGA - LOHMKRRDDNNHEVIGT - QKNSTKENEEVYEVVGA	TSSHYADLNPQFITPSVYAE       419         TSSHYADLNVKSVKPSIYAD       441         TLSNYADLNVQAIKPSLYTD       434         TSHYADINIKAVKPSLYTD       428         TSSHYPDQNVKYVKPCTNAD       428         TVSNYTDLNLKAVKPVKPCTNAD       428         TVSNYTDLNLKAVKVVKPCTNAD       428         TVSNYTDLNLKAVKVVKPCTNAD       428         TVSNYTDLNLKAVKPVKPCTNAD       428         TVSNYTDLNLKAVKPVKPCTNAD       428
	Alr16 Alr18 Alr19 Alr21 Alr36	442         LTKATD PTKKYAE           435         LTTATR SSKHYTE           429         LTTTAD HSKQYAE           429         LTTTTD SSKGYAE	I <mark>E</mark> LH <mark>K</mark> KDT <mark>Y</mark> SNIRD IEFSETKQSMHANIST ADAYKSDENMYEM	453 467 463 455 457 451 782
С		373 KYRGAETNCKNKQSAK 375 KYRGAETNCKNKQSPK	KDDDNKDAQDYEVFPNAVS NDDENEDRQEYEVFPTTES NHDENKDRQEYEVFPTTES KDDEYKDSQDYEVFPATES	HYTGLQLEA <mark>RK</mark> EIVYADLL 426 HYTGLQLEA <mark>RKEIVYADL</mark> S 428
	Alr11 Alr12A Alr12B Alr9	423 PTPVNEYAEVGTENNS 427 PPTVNNYSEIGTEKRS 429 PPTVNNYSEIGTEERS 428 PATVNEYSEIKTENQL	KQF KQF	441 445 447 446
D	Alr1.1 Alr4 Alr6.1 Alr1.4 Alr6.2	382 KRBRQKKQDAADGFRMI 362 KRBK NDESNYPLF 184 RWKKNH PSKTDTSEC	L <mark>SP</mark> HQSENEYATPVLLE RYEEPEENDYVMDLNLPKTP QMKSLKTTK <mark>S</mark> TEEGVHY	NGRPPAN PDDPEQAIYSE 432 DSQQNEE EDPNHALYSK 430 NAOPSSNEDRKNNPDQVIYSD 413 YASTDVTVP-PEKRDQQVYAQ 233 YA <mark>QP</mark> SAAM SQDQPKQVYAE 226
	Alr4 Alr6.1 Alr1.4	431 L GPGGGRTGPRPALE 414 V GMGGGRTGPKPVV 234 VDRSGGGGRDGPKPEV	E F <mark>S</mark> N Y A E I <mark>K</mark> VD AMG Y P VD G / A P A I Y SEMK VD S RG Y P VD G / K <mark>SD Y AQ</mark> MQ VD ADG Y P A SG I	AKA SE <mark>PP</mark> TYA <mark>PIIKPREG</mark> AKR 486 AKE I A SHADYA SI DD 478 NNP SREKGVY SAV <mark>GP</mark> 461 PT S
	Alr4 Alr6.1 Alr1.4	479 CMM <mark>SPP</mark> SDDDDDA 462N SH SD SD SD A 272 <b>S</b> T PPD EN EDG I S		SRSPPPDEDDHEGV       IV       537         IV       IV       494         IV       PHYAGV       VV         SGV       VV       288         RGIPRNSAE       285

# Fig. S6

- (A) Alr cytoplasmic tails grouped by CD-HIT at 20% similarity. Neighbor-joining trees are shown. Leaves are color coded according to their genomic position. Branch lengths calculated according to the BLOSUM26 matrix. Scale bar = 100 units.
- (B) Alignment of group 1 cytoplasmic tails.
- (C) Alignment of group 2 cytoplasmic tails.
- (D) Alignment of group 3 cytoplasmic tails.

#### Α

A			
Alr16 390	-KHKGFIIFIKNPKKRLQHTKRRNDDNYEVTQI	TSSH <mark>YADLNVKSV</mark>	<pre>KPSIYADLTKATDPTKKYAEIELSEIKRNMHAN 467</pre>
			KPCTNADLTTTTDSSKGYAEIEFSETKQSMHANIST 457
Alr36 377	-KRKVLKEIFKEKVNHQKNSTKENEEVYEVVGA	TVSN <mark>YTDLNLDKT</mark>	VTNLYTGLTKREKENTAADAYKSDENMYEM 451
			KPSVY <mark>ADL</mark> KTTTDDSNQYAEIELNDVK 782
			TPSVYAELTNNTTAEAASSNYAEPNVQAIKPRLYADLTTKC 453
Alr18 383	-KRKGLVTVNIRRKNQSKTDVKDNIDD	TLSN <mark>YADLNVQAI</mark>	KPSLYTDLTTATRSSKHYTEIELKETQKNTYANIKT 463
Alr19 378	-KHKGYITVSINRKRRPQSNENGS-DDYEETGA	TTSH <mark>YADINIKAV</mark>	<mark>KPSIY<mark>ADL</mark>TTTADHSKQ</mark> YAEIELHKKDT <mark>Y</mark> SNIRD <i>455</i>
-			
В			
	KLRGTGCNCRNKQSVKKDDE <mark>Y</mark> KDSQD <mark>Y</mark> EVFPATESI		
	KLRGSQTNCRNKESVKKDDDNKDAQD <mark>Y</mark> EVFPNAVSI		
Alr12A 373	KYRGAETNCKNKQSAKNDDENEDRQEYEVFPTTESI	HYTGLQLEARKEI	VYADLLPPTVNNYSEIGTEKRSKQF 445
Alr12B 3/5	KYRGAETNCKNKQSPKNHDENKDRQEYEVFPTTESI	HYTGLQLEARKEI	VYADLSPPTVNNYSEIGTEERSKQF 44/
С			
-	SVTIKELEENENVEENND <mark>YKDLNCLDISEET</mark> YSKL	TPPKSRLHKARVO	LRYSKGKIKSFKKHSTKSTHEYEVPPSLLANMSKDKTAD 463
D			
_	Syk		Shark
huma	n • 🚰 💳 🚰 💳 🖉 🛶 🖉	drosophila	
Hydractini		Hydractinia	
Tyaraolini		riyaraolinia	
	l + + + + + + + + + + + + + + + + + + − + □ = 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0 200 400 600 800 aa
	0 200 400 600 aa		0 200 400 800 aa
	SH2 kina	ase	ankyrin repeat
_			
E	SHP		SHIP
	SHP-1		SHIP1
human		humon	Shipi
numan	SHP-2	human	SHIP2
			SHIF2
		Hydractinia	
		-	
			0 200 400 600 800 1000 1200 aa
Hydractinia	a		
-			
			SH2 phosphatase SAM

# Figure S7. ITAMs in putative AIr proteins

- (A) Alignment of group 1 cytoplasmic tails, showing ITAMs (orange shading). Overlapping ITAMs are shown with heavier shading. Bona fide *Alr* genes (blue gne names) are included in the alignment for comparison.
- (B) Alignment of group 2 cytoplasmic tails, showing ITAMs. Shading as in (A).
- (C) ITAM in putative Alr gene Alr31.
- (D) Comparison of human Syk and *Drosophila* Shark to *Hydractinia* Syk and Shark.
- (E) Comparison of human SHP and SHIP proteins to *Hydractinia* SHP and SHIP homologs.

<i>Hydractinia</i> Syk <i>Hydra</i> Syk Human Syk <i>Hydractinia</i> Shark <i>Hydra</i> Shark Drosophila Shark	MADPASLPWFHGRITRDTAEKCLHK-TSMEDGTYLLRESTSQIGSYVLSVCKDRKVVHYQIQKQPSGMVGIADGPKPGPVELVNHHKTNLDGLMTKLSVACNRLAGVAKA 111 MADPSLLPWFHGRITRDIAENCLQTTSMDDGTYLLRESTSEIGSYVLSVCKDKKVIHYQIQKQPSGMVGIADGPVFLVNHHKTNLDGLMTKLSVACNRLAGVAKA 111 MASSGMADSANHLPFFFGNITREAEDYLVQ-GGMSDGLYLLROSRNYLGGFALSVAHGRKAHHYTIERELNGTYAIAGGPVFLVNHHKTSLGGLMTKLTYPCEPAGVKATA 119 MTYNNGSLLWYHGKISREVATQTLSRNGTGRDGCYLIRDCSSAPGDYVLSLWNKNQVMHFQVHCLGDNKFSIDDGPIFNGLDSLTAHYRTNPDGLPCKLTSFCQGRMPPLNTLKYGKETK 120 MTYNNGSLLWYHGKISREVATQTLSRNGTGRDGCYLIRDCSSAPGDYVLSLWNKNQVMHFQVHCLGDNKFSIDDGPIFNGLDSLTAHYRTNPDGLPCKLTSFCQGRMPPLNTLKYGKETK 120 MSKNSDÄLLWYHGKISREVATQVLRKG-GRDGFFLIRDCGNAPEDYVLSLWNKNQVMHFQVHCLGDNKFSIDDGPIFNGLDSLTAHYRTNPDGLPCKLTSFCQGRMPPLNTLKYGKETK 120 MSRDSDPMKWYHGNLSREAADELLKQ-G-YEDGTFLIRDCGNAPEDYVLSMMFRSQILHFQINCLGDNKFSIDDGPIFNGLDSLTAHYRTNPDGLPCKLTSFCQGRMPPLNTLKYGLDTR 119 MSRDSDPMKWYHGNLSREAADELLKQ-G-YEDGTFLVRESSTAAGDFVLSLLCQGEVCHYQVRRHGGEDAFFSIDDKVQTKILHGLDTLVDYYQQAANGLPTKLTVPLIRDLPPHNTRSHGVTNL
<i>Hydractinia</i> Syk <i>Hydra</i> Syk Human Syk <i>Hydractinia</i> Shark <i>Hydra</i> Shark Drosophila Shark	YNDYSHDDLDEA-TRVALRDI149 
Hydractinia Syk Hydra Syk Human Syk Hydractinia Shark Hydra Shark Drosophila Shark	C-SH2 domain VVGNILHKKQPWFHGAIPRNEADRRLETHNFEGMYLIRERGSY-RNSYVLGLCHQKKVYHYLFEPNDKGQLSIKAGRPF 228 
Hydractinia Syk Hydra Syk Human Syk Hydractinia Shark Hydra Shark Drosophila Shark	C-SH2 domain
Hydractinia Syk Hydra Syk Human Syk Hydractinia Shark Hydra Shark Drosophila Shark	TKPIVP-TRPPPSPAEQPVA-Q
Hydractinia Syk Hydra Syk Human Syk Hydractinia Shark Hydra Shark Drosophila Shark	PTKc domain G-FNSPFQKNKKRDISLDPSRLSL-GSDLGEGNFGSVKKGTYVLPKGNQILVAVKTLKEDIPGQKSEILSEANIMAHLDHPNIVRLIGVTQSPNFMLVMELAPEGPLLGYLKKHRSMPML 457 G-FKNPFETQERETELDPKRLRL-EHELGHGNFGSVVKGVYLLKNNKSVAVAVKTLKEDIPGQKSEIVKEAEIMSKLDHPNIVRLIGVTQSPNFMLVMELAPQGPLHKFLKKNRMNVL 455 V-YESPYADPEEIRPKEVYLDRKLLTLEDKELGSGNFGTVKKGYYQMKKVVKTVAVKILKNEANDPALKDELLAEANVMQQLDNPYIVRMIGICEAESWMLVMEMAELGPLNKYLQQNRVKDK 469 P-VAKQQSSVQQTMIPRESLEL-GQELGVGEFGSVLKGVWTSPDGEKISVALKTLHQDKLQQEKEFLREARVMSQLNHPCIVRLLGVCLGPPMILVQELITLGALLDHLIDHOPEIQ-EV 587 G-DQTMKNNAQQNIILKESISF-GKELGVGEFGSVLKGIWLSPGGKEINVAMKTLHKDKMVQGEKEFLREALVMSQLNHPCIVSLLGVCLGPPMILVQELVEMGALLDYLMDYQPEIQ-EV 583 GTPSTPSATEVEAAKLRFFIEPEKLVL-DREIGHGEFGSVHSGWLLRKSGAGEESRLEVAIKMLSDEHSNKQEFLREASVMMRLEHKCIVRLIGIAKGEMLMMVQELAPLGSMLQVILDHGHEITANA 764
<i>Hydractinia</i> Syk <i>Hydra</i> Syk Human Syk <i>Hydractinia</i> Shark <i>Hydra</i> Shark Drosophila Shark	PTKc domain NVLILMLQVDEGMNYLESQHFVHRDLAARNILVVSENFVKISDFGMSRAMGAGNEYYRAERTGKWPLKWYAPECIYYHKFTSKGDVWSYGVALWEAVSYGMKPYQGMKGPQIVEMLDNGQRLSRPERCPE 587 DILVLMLQVDEGMNYLESQHFVHRDLAARNVLVVSENFVKISDFGMSRAMGVGNEYYRAERAGKWPLKWYAPECIYYHKFTSKGDVWSYGIIMWEAVSYGGKPYQGMKGPQIVEMLDNGQRLSRPERCPE 587 NIIELVHQVSMGMKYLEESNFVHRDLAARNVLVVTGHYAKISDFGLSKALRADENYYKAQTHGKWPVKWYAPECINYYKFSSKSDVWSFGVLMWEAFSYGQKPYRGMKGSEVTAMLEKGERMGCPAGCPR 599 DLKLWAAQIASGMMYLEHRRFVHRDLAARNVLLVTGHYAKISDFGLSKALRADENYYKAQTHGKWPVKWYAPESINYGTFSQKSDVWSYGVLWEMFFFGELPYGEMTGAEVIALLERGQRLEKPDDCPE 717 DLKLWASQIAFGMMYLEHRRFVHRDLAARNILLANKKQVKISDFGLSRAVGAGSDYYQAQQGGRWPVKWYAPESINYGTFSQKSDVWSYGVLWEMFFFGELPYGEMTGAEVIALLERGQRLEKPDDCPE 713 DLKLWASQIAFGMMYLELKRFVHRDLAARNILLANKKQVKISDFGLSRAVGTGSDYYQAKQGGRWPVRWYAPESINYGTFSTKSDVWSYGITLWEMFFFGDLPYGEMIGNEVVSFLEHCGRLEKPDECPI 713 ELKVWASQIACGMHYLESQHFVHRDLAARNILLANKKQVKISDFGLSRAVGTGSDYYQAKQGGRWPVRWYAPESINYGTFSTKSDVWSYGITLWEMFFLGDLPYGEMIGNEVVSFLEHCGRLEKPDECPI 713 84000000000000000000000000000000000000
Hydractinia Syk Hydra Syk Human Syk Hydractinia Shark Hydra Shark Drosophila Shark	PTKc domain

**Figure S8. Syk and Shark protein alignment.** ClustalO alignment of *Hydractinia* Syk-like and Shark-like proteins with *Hydra* Syk-like (NP\_001296711), Human Syk (NP\_001167638), *Hydra* Shark-like (NP\_001296681), and Drosophila Shark (NP\_524743). Ankryin repeats are boxed in black.

Hvdractinia SHP-like 3	MVRWFHRDLSGLDAETLLKGRGVHGSFLARPSRKNQGDFSLSVRVGDQVTHIRIQNSGDFYDLYGGEKFATLTELVEYYTQQQGVLQDRDGTIIHLKY MTSRRWFHPNITGVEAENLLLTRGVDGSFLARPSKSNPGDFTLSVRRNGAVTHIKIQNTGDYYDLYGGEKFATLAELVQYYMEHHGQLKEKNGDVIELKY MRSSWGFHAQLTFDEAEKILKEKGANGNYLCRRSATTPGSYTLSVKRGNVVWHFKIRNDGDCFELYENDGFASVPDLIEYYQQNPSKFIDADGNCVQMSE MRPTWGFHPNITLEGAIQLLLEKGKHGSYICRRSEKVQDEYRLSVKRGNVVWHFKIRNDGDCFELYENDGFASVPDLIEYYQQNPSKFIDADGNCVQMSE MRPTWGFHPNITLEGAIQLLLEKGKHGSYICRRSEKVQDEYRLSVKRGNVSHFKIRNDGDCFELYENDGFASVPDLIEYYQQNPSKFIDADGNCVQMSE MRPTWGFHPNITLEGAIQLLLEKGKHGSYICRRSEKVQDEYRLSVRKNONSVTHIKIQNHGDYYDLPGFSTLSELIQYYMERELREKNGELIELLY	100 100 100 19
Hydractinia SHP-like 2 Hydractinia SHP-like 3	C-SH2 domain PLNCSDPTSERWYHGHMSGGQAETLLQAKGEPWTFLVRESLSQPGDFVLSVLSDQPKAG-PGSPLRVTHIKVMCEGGRYTVGGLETFDSLTDL PLNCADPTSERWFHGHLSGKEAEKLLTEKGKHGSFLVRESQSHPGDFVLSVRTGDDKGESNDGKSKVTHVMIRCQELKYDVGGGERFDSLTDL PVAYDDDSDPGLDKERWFFGEIGRGEAHSFLKQRGEDGSYLVRESRSKPGSYVLSIRHKNEVTEFLIEYTDGTFDVSGTKSLKFPTMQHL PLPSEDLLKLGLLNERWYAGYISMKETTSLLKQEGESGSYFIRESSEPGSYALAVRLDGAVIDFIIRYKDGRFHIREE-FVYFTTIQSL PVRSNDPTTERWYHGPISSKEAERMLLAKGRQGSYLVRESQSQPGQYALAVRCNNGIQQIIIRHRETKYDIGSGPTFLSLREL PLNCKDPTSERWFHGHMSSKVAERLMLDKGKNGSYLVRESVSKPGDYVLTIKSDETVLHVMIYNVDGKYDIGGGPVFETITDL	193 190 189 102
Human SHP-1 Human SHP-2 Hydractinia SHP-like 1 Hydractinia SHP-like 2 Hydractinia SHP-like 3 Hydractinia SHP-like 4	C-SH2 domain VEHFKKTGIEEAS GAFVYLRQPYYATRVNAADIENRVLELNKKQESEDTAKAGFWEEFESLQKQEV - KNLHQRLEGQRPENKGKNRYKNILPFDHSRV VEHYKKNPMVETL GTVLQLKQPLNTTRINAAEIESRVRELSKLAETTDKVKQGFWEEFETLQQQEC - KLLYSRKEGQRQENKNKNRYKNILPFDHTRV LNHFTAHPPMDVH GNVVPLKEALCDTAER R KNDQ RMKKEFEWLQQEDR - ANTNTKSEG LKPENRGKNRYKNILPFDHTRV LKYYVHNDLFDVN GCVVPLKEPLFDTIEK R KTEKELQKIQSEFEALQEGDR - TNTYSRDEG LMPENIKKNRYKSILPFDHTRV VEYYVPNKLHDIKDQSPIKLREPFQSTSFIASSVEERMEVLMKENP KGATGFREEFEQLQRFDHNDGVYTCSDG0TEENKWKNRYKNILPYDHSRV VDYYRQNPMVEKN - GGTVVHLKQPFNATKITASSIGDRVVELQRETP - LVPNKDGFWEEFEQLQRPEC - KHRYSRKVGLSSENKIKNRFKNIVPFDHTRV	287 290 269 271 199
Human SHP-1 Human SHP-2 <i>Hydractinia</i> SHP-like 1 <i>Hydractinia</i> SHP-like 2 <i>Hydractinia</i> SHP-like 3 <i>Hydractinia</i> SHP-like 4	Tyrosine Phosphotase domain ILQGRDSNIPGSDYINANYIKNQLLGPDENAKTYIASQGCLEATVNDFWQMAWQENSRVIVMTTREVEKGRNKCVPYWPEVGMQRAYGPYSVTNCGE VLHDGDPNEPVSDYINANIIMPEFETKCNNSKPKKSYIATQGCLQNTVNDFWRMVFQENSRVIVMTTKEVERGKSKCVKYWPDEYALKEYGVMRVRVKE VLRNVDPNISGSDYINASYIFDDESGALFIASQGCVKATVNDFWHMIDQENSRIIIMUTNEVEKAKIKCVRYWPEPGVSMNVKGKIVTNTGE LLQNIDPNTSGSDYINASYISDDKSGLVFIATQGCLKTTFNDFWHMINQENSGVIIMLVKEIEAEKRKCVKYWPETSLSDVDGLVVKNVNE RLKGGNPKDIGRDYINANYIDVELQSWKVKYIASQGCLPSTIGDFWKMIYQHNSSIVMLTNEVEKAKSKCAKYWPNLGQSVEYDKIVVDNMLE ELLDGDPDDPTQDYINASYVKGESDHEYIATQGPLAETIHDFWRMVYQENSCVILMITREVERGKSRCAKYWPDVDLSMEVENFTIFNSKE	390 361 363 293
Human SHP-1 Human SHP-2 <i>Hydractinia</i> SHP-like 1 <i>Hydractinia</i> SHP-like 2 <i>Hydractinia</i> SHP-like 3 <i>Hydractinia</i> SHP-like 4	Tyrosine Phosphotase domain HDTTEYKLRTLQVSPLDN-GDLIREIWHYQYLSWPDHGVPSEPGGVLSFLDQINQRQESLPHAGPIIVHCSAGIGRTGTIIVIDMLMENISTKGLDCD SAAHDYTLRELKLSKVG - Q-GNTERTVWQYHFRTWPDHGVPSDPGGVLDFLEEVHHKQESIMDAGPVVVHCSAGIGRTGTFIVIDILIDIIDEKGVDCD TATRDYVVRQLEISDELHGNQPTPPRKIFLYQFVGWPDHGVPGDPGSLLEVMQLIEATQKTFEYPGPPVIHCSAGIGRTGTVIVISMLMSLYRIKGHLEE TVTQDYIMRQLEIIEDKV SSTKTIHHFQFIAWPEHGVPSDPGSLLEFIQVEITRRRTHCDGPPVVHCSAGIGRTGIVVVISMLMHLYQMKGMLED HKTKTFTMRELRIYHSD KAKKAPLMVYQYHFTDWPDHGAPDDPGAILGLLNEIHTKHDSCGRNAPIVVHCSAGIGRTGTVMVIDVLMHLLDEKGLNTE TETQEFIYRELKLTNGD EPEEGEHTIYHYQYIGWPEHGTPSNVGSVIGILHDINLKQKSNNYPGPIVAHCSSGIGRTAAFLVIDILVKILQKQQLDCE	487 461 459 391
Human SHP-1 Human SHP-2 Hydractinia SHP-like 1 Hydractinia SHP-like 2 Hydractinia SHP-like 3 Hydractinia SHP-like 4	Tyrosine Phosphotase domain IDIQKTIQMVRAQRSGMVQTEAQYKFIYVAIAQFIETTKKKKLEVLQSQKGQESEYGNITYPPAMKNAHAKASRT IDVPKTIQMVRSQRSGMVQTEAQYRFIYMAVQHYIETLQRRIEEEQKSKRKGHEYTNIKYSLADQTSGDQSPLPP RDVPRTVQKVRLQRSGMVQTEAQYRFIYMAVVFFMETVERREVGAQQSRNLYGNLDAMSELLSDLTGDQEDLPPPPPRST EAIFQTVRNVRLQRLGMVQTKVQYGFVYKAMKHFMQTSAQKDDAESKETKNYDNLDNHVSEICRNQLWWTRLRDEE IDIKKTAQLLRTQRSGMVQTEMQYRFVYQAIKHFIDTEGKRIAAKKDMTTSGRPSYGNIDFSKP	562 541 536 469
Hydractinia SHP-like 2 Hydractinia SHP-like 3	CSSKHKED-VYEN-LHTKNKREEKVKKQRSADKEKSKGSLKRK 595 CTPTPPCAEMREDSAR-VYEN-VGLMQQQKSFR593 KASMKSKPVPKPN554 	

**Figure S9. SHP protein alignment.** ClustalO alignment of *Hydractinia* SHP-like proteins with Human SHP-1 (NP\_002822),and SHP-2 (NP\_002825)

Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	MYPCWNHGNITRSKAEELLSRIGKDGSFLVRASESISRAYALCVUYRNLPNEDDKFTVQASEGVSMRFFTKLDQLIEFYKKENMGLVIHLQYPVPLEEDTGDDPEE-DIVESVVSPPELPPRNIPL 133 MASACGAPGPGGALGSQAPSWYHRDLSRAAAEELLARAGRDGSFLVRDSESVAGAFALCVLYQKHVHTYRILPDGEDFLAVQTSQGVPVRFQTLGELIGLYAQPNQGLVCALLLPVEGEREPDPPDDRDASDGEDEKPPLPPRSGST 148 
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	TASSCEAKEVPFSNENPRATETSRPSLSETLFØRLQSMOTSØLPEEHLKAIQDYLSTQLAQDSEFVKTGSSSLPHLKKLTTLLCKELYGEVIRTLPSLESLQRLFDQQLSPGLRPRPQVPG-EANPINMVSKLSQLTSLLSSIED 277 SISAPTGPSSPLPAPETPTAPAAESAPNGLSTVSHDYLKGSYGLDLEAVRGGASHLPHLTRTLATSCRRLHSEVDKVLSGLEILSKVFDQQSSPMVTRLLQQNLPQTG-EQELESLVLKLSVLKDFLSGIQK 280 SITLFFQQQLDIIKGLKLDTVLKESLQEYVTKGINKDFØRVQSGSNKPPMLQKLLVEAAGDLQLSLQKFLTRLGHIHELFDVGSNVQALEQYEHINQDNNSKDLESVFKMLSECKSGMLRLEQ 244
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	KVKALLHE <mark>GP</mark> ESPHRPSLIPPVTFEVKAESLGIPQKMQLKVDVESGKLIIKKSKDGSEDKFYSHKKILQLIKSQKFLNKLVILVETEKEKILRKEYVFADSKKREGFCQLL 388 KALKALQDMSSTAPPAPQPSTRKAKTIPVQAFEVKLDVTLGDLTKIGKSQKFTLSVDVEGGRLVLLRRQRDSQEDWTTFTHDRIRQLIKSQKPQNKLGVVFEKEKORTQRKDFIFVSARKREAFCQLL 408 KACKVLKEYSGLSTSQNEQK <mark>P</mark> SNTEMEYSVLSYDTKGSPTVIAEKPKDVKHKLFEVTSGPLKSKSYIALDITNGKIIFLKNNQDKLESGTAFTSEQVFQLVKSKTSKSKLGILLENMNVKNYYFENFRSRELFCQVV 381
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	Phosphotase domain QQMKNKHSEQPEPDMITIFIGTWNMGNAPPPKKITSWFLSKGQGKTRDDSADYIPHDIYVIGTQEDPLSEKEWLEILKHSLQEITSVTFTVAIHTLWNIRIVVLAKPEHENRISHICTDNVKTGIANTLGNKGAVGVSFMFNGTSLGFV QLMKNKHSKQDEPDMISVFIGTWNMGSVPPPKNVTSWFTSKGLGKTLDEVTVTIPHDIYVFGTQENSVGDREWLDLLRGGLKELTDLDVRPIAMQSLWNIKVAVLVKPEHENRISHICTDNVKTGIANTLGNKGAVGVSFMFNGTSFGFV 538 QQIKIAHSKDIHSSELSVFVGSWNMGDAVE-GDISEWFKCLGLGKTRPQEFCHIASDIYAIGTQESAYHEKEWVKKIQTCLKSVFRKEFKLLTCYSMWDIRIVIFVKSELLNLVNSVKQSSVKTGIATVLGNKGAVGVSFNVGTSFCFV 530
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	Phosphotase domain NSHLTSGSEKKLRRNQNYMNILRFLALGDKKLSPFNITHRFTHLFWFGDLNYRVDLPTWEAETIIQKIKKQQQYADLLSHDQLLTERREQKVFLHFEEEEITFAPTYRFERLTRDKYAYTKQKATGMKYNLPSWCDRVLWKSYPLVHVVCQ NCHLTSGNEKTARRNQNYLDILRLSLGDRQLNAFDISLRFTHLFWFGDLNYRLDMDIQEILNYISRKEFEPLLRVDQLNLEREKHKVFLRFSEEEISFPPTYRYERGSRDTYAWHKQKPTGVRTNVPSWCDRVLWKSYPETHIICN NCHLTSGNEKLQRRNQNFHSILRGLNLGQRDSFDLMLSFHHLFWFGDLNYRLDLDVTDIIKWIMEREWQKLLAADQLIAEIRKGNAFISFEEDISFPPTYRYTRGTRSTYCWTKEKRTGTRINVPSWCDRVLWKSYPELTSCN 675
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	GCIALRLEATETQLPIYTPLTHHGELTGHFQGEIKLQTSQGKTREKLYDFVKTERDESSGPKTLKSLTSHDPMKQWEVTSRAPPCSGSSITETINPNYMGVGPFGPPMPLHVKQTLSPDQQPTAWSYDQPPKDSPLGPCRGESPP 962 CVVALKSMIGSTAQQFLTFLSHRGEETGNIRGSMKVRVPTERLGTRERLYEWISIDKDEAGAKSKAPSVSRGSQEPRSGSRKPAFT CVISMKMFLSGNPKVFEVMITHLGVETGVLKVEAQLIIPSTIEYTASQKSKVYDLIKVEEEDSRPNTKKC-GEKKEMYLVNLPVKKEDIDQIDPVDKKVEPP 926
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	TPPGQPPISPKKFLPSTANRGLPPRTQESRPSDLGKNAGDTLPQEDLPLTKPEMFENPLYGSLSSFPKPAPRKDQESPKMPRKEPPPCPEPGILSPSIVLTKA 1065 APPRAAPRE-EPLTPRLKPEGAPE
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	QEADRGEGPGKQVPAPRLRSFTCSSSAEGRAAGGDKSQGKPKTPVSSQAPVPAKRPIKPSRSEINQQTPPTPTPRPPLPVKSP-AVLHLQHSKGRDYRDNTELPHHGKHRPEEGPPGP
Human SHIP-1 Human SHIP-2 <i>Hydractinia</i> SHIP-like	SAM motif RIRESIQEDLAEEAPCLQ <mark>GG</mark> RAS <mark>GLGEAGMS</mark> AWLRAIGLERYEEGLVHNGWDDLEFLSDITEEDLEEAGVQDPAHKRLLLDTL-QLSK 1189 RTKKMVKQFVSDFSKWLNNIGFEMYAEHLINNGFDDLDYFCSITEEDLKDVGILDEKHREKIVLEAKKLKGRHKDHKDIVAV 1141

**Figure S10. SHIP protein alignment.** ClustalO alignment of *Hydractinia* SHIP-like protein with Human SHIP-1 (NP\_001017915),and SHIP-2 (NP\_001558)

Alr1 Domain 1  A-   A'   -B   C'-   C''   -D   E-   F   F   G  L <mark>NVKF</mark> VLP <mark>TIST</mark> TFNAT <mark>VELQWQVT</mark> LEPSEAIA <mark>SMTVREI</mark> PSL <mark>ITILTGA</mark> VD <mark>GFNV</mark> A <mark>QGGRKLF</mark> GDR <mark>VSGIFN</mark> NRNNIVTETL <mark>TFQLLVG</mark> SSTL <mark>VKAANISIVEI</mark> S EEEETTTEEEEETTT EEEEEEEETTTEEEEEEETTEEEEE HHHHHHHTTTEEEEEEETTTTEEEEEEETTTTEEEEEEETTTTEEEE
Alr2 Domain 1 A  A'   B   -C   -C'-  C"  D   -E-   F   G  LS <mark>LT</mark> SPA <mark>TIEE</mark> VVGRS <mark>VTITYVT</mark> DVADVADVVDV <mark>VNFRINY</mark> ND <mark>SRIAEG</mark> TKTV <mark>FDK</mark> IPPTPFGNR <mark>LRT</mark> STPLQNQK <mark>EYSLL</mark> DNLEYNDT <mark>GLFFAKIET</mark> FKPN <mark>VEANSTTNLLT</mark> Y EE TTEEEETTT5 EEEEEEEE TTTTTTT EEEEEEETTTEEEEETTTTEEEETTTTTT
Alr3 Domain 1  A   A'   B   C   -C'-   C"   -D   -E   F   G  A <mark>VLY</mark> TDSP <mark>TISG</mark> IFNQS <mark>AIISFKAT</mark> ITANDN <mark>RTVKSFYVE</mark> LPPSTQ <mark>QIAIGS</mark> NN <mark>VLQA</mark> LPVPPFNSR <mark>.TASAA</mark> NQ <mark>EYLLQV</mark> NPLKFSDD <mark>YTFRGVLIYLL</mark> NNDINN <mark>PIAVEQDVKLDV</mark> F EEETTTTEEEEETTTD EEEEEEEETTTTT EEEEEETTEEEETTTTTTTEEEEEE
Alr4 Domain 1        A   A'    B   C'-  C''  -D   E   F   G          NVEPLNGQSEVTGMLDKDITLQWQ1TFLKGEMLQSHDIY1PNRTKIVSNQPPELTPVGKRMYGTRLVPVFDADAAVFKLLKNVKFTDSSHNFTLVVAFERKDDFNRTGVADINIVNVE         EEEEGGG       EEEETTTEEEEEEEETTTEEEEEETTTEEEEETTTEEE
Alr6 Domain 1  A-   A'   B   C   -C'  C"  -D-   -E   F   G  GTVEAVKTNFEV EEEETTTEEEETTT EEEEEEEE TTTEEEEEEETTTEEEEEE
Alr7 Domain 1  -A   A'-   B   C   -C'-   C''   -D-   -E   F   G  ASSG <mark>KVTV</mark> VKN <mark>VFERV</mark> QIKST <mark>VVMEWKIA</mark> HSDN <mark>QTNKVLKLYVL</mark> PDRDR <mark>PVFSSY</mark> GK <u>VQSS</u> QCKGRQTFGNRLSATFSKIKG <mark>KYTVIL</mark> KRIQYNEN <mark>YTFQLKVIFIN</mark> KKSV <mark>TETKVADIRIKNV</mark> V EEEETTTEEEEETTT5 EEEEEEEE TTTEEEEEEEETTTTEEEEEETTEEEE HHHHHHHTTTEEEEEEGGG EEEEETT5TTT EEEEEEEETT EEEEEEEEEE
Alr8 Domain 1a  A   A'   -B   C   C'-  C"  -D-   -E-   F   G  YGG <mark>TITA</mark> AAT <mark>VVTV</mark> NFQSTLKLEWTAIPSPAEQIVVLKVFILPDTNNGVLTNTNPPVLLPTGMSLFGRNRLSATFINSKYTLMLKNVTYNDSCTFQMYAIFRKPAAYSV EEEETTTEEEEETTT EEEEEEEE TTTTEEEEEEEETTTTEEEEETTTEEE HHHHHHHTTTTEEEEEETTEEEEEETTb GGG EEEEEEEETTTTTTEEEEEEEEE
Alr8 Domain 1b  -A-   A'   B   C   C'  C"  -D   -E   F   G  VNDETISDTASTLDIVYNTTLKLDWNMTLSAGEVIVGVQVVLPDTTNRIITDTSPTVLSKGISIFGENRLSATFTNSRYRLMLRNIRYNESFTFQLVVIYGAGDHLSSSRFN1QVTV EEEEE TTEEEETTTDEEEEEEE TTTTEEEEEETTTTEE HHHHHHHTTTTEEEEEETTBETEEEEETTBTTTEEEEEETTBEEEEEE
Alr9 Domain 1  A-   A'   B   C   -C'  C"  D-   E-   F   G  G <mark>SVEA</mark> IIK <mark>NIEA</mark> TDNST <mark>AELSWRVE</mark> TNKDGE <mark>RVFGVELSE</mark> GG <mark>VVVID</mark> DKSG <mark>TVTEAGRNKF</mark> GGR <mark>LSASFS</mark> NN <mark>VYKMFI</mark> KKIQYNEAK <mark>SFTLTAAFYW</mark> KS <mark>ALDPVNDTATITSV</mark> K
EEEETTTEEEEETTTb EEEEEEEE TTTTEEEEEEETTEEEEETTTEE HHHHHHHH
EEEETTT EEETTTb EEEEEEEE TTTTEEEEEEEEETTTTEEEEETTTEE HHHHHHHH
EEEETTTb EEEEEE TTTTEEEEEEEETTTTEEEEETTTEE HHHHHHHH
EEEEEETTTb EEEEEE TTTTEEEEEEEETTTTTEEEEEEE HHHHHHHH
EEEETTT EETTTE EEEEEETGGGEEEEEETTEEEEEE GGGTTTEEEEEETTEEEEEEETTEEEEEEETE GGGTTEEEEEEEE
EEEETTTEEEETTTb EEEEEEEE TTTTEEEEEEEETTTEEEE HHHHHHHHTTTEEEEEEGGG EEEEETTbTTTT EEEEEEEEETTEEEEEEEEEE
EEEETTTEEEEETTTE EEEEEEE TTTTEEEEEEEETTTTTEEEEEE
EEEETTTEEEETTTb EEEEEEEE TTTT EEEEEEETTTEEEEEETTEEEE HHHHHHHH
EEEETTTEEEEETTTE       EEEEEEE       TTTT       EEEEEETTTEEEEEEETTTEEEEEE       HHHHHHHHTTEEEEEETTTEEEEEETTTEEEEEETTTEEEEEE
AVELIAR RECOVERY AND SEPTEMBED SEPTEMBED SEPTEMBED SECTION AND AND AND AND AND AND AND AND AND AN
EEEEETTTEEEETTTE EEEEEEETTTTEEEEEEEEEE
PERSONARCAL STRACT DE DOMARTA DE DOMARTA DE DOMARTA DE LA RESTOTET DE CONTROLLE ANVESTED DA DE DOMARTA DE DOMART
IFRAECAEON <mark>LIGRAERGNNVILIWS</mark> ENVIRRI <mark>IIRTEIRF</mark> DNIVVSICNAAGENG <mark>TIUV</mark> IFRNDA <mark>IEASEN</mark> KIDK <mark>UILLI</mark> ENETISDS <del>DIEVADDIAM</del> GA <u>U</u> RPERLINTIIOVR EEEEETTTE EEEEEE TTTT EEEEEEEETTEEEEEEEE

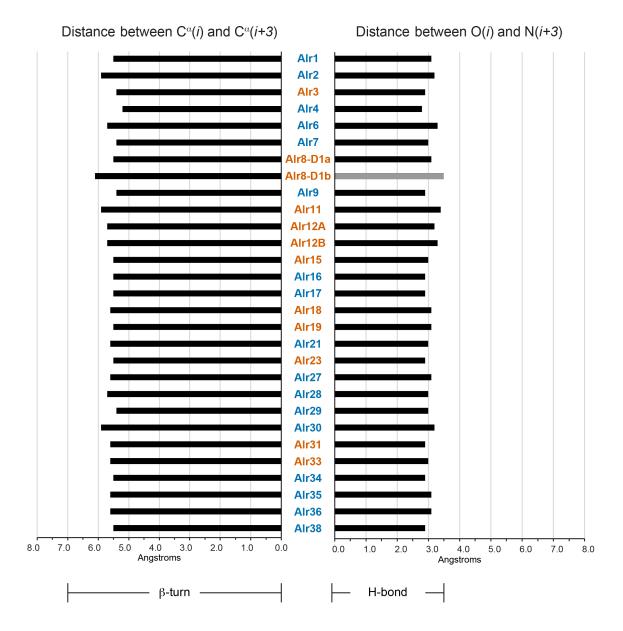


# Fig. S11. STRIDE secondary structure predictions for AIr domain 1

For each domain, the top line shows beta-strands labeled according to their position in the primary amino acid sequence. The middle line shows the sequence of the domain. The bottom line shows the STRIDE secondary structure predicted from the Colabfold model. (H = alpha helix, G = 3-10 helix, I = PI-helix, E = beta-strand extended conformation, B = isolated bridge, T = turn.)

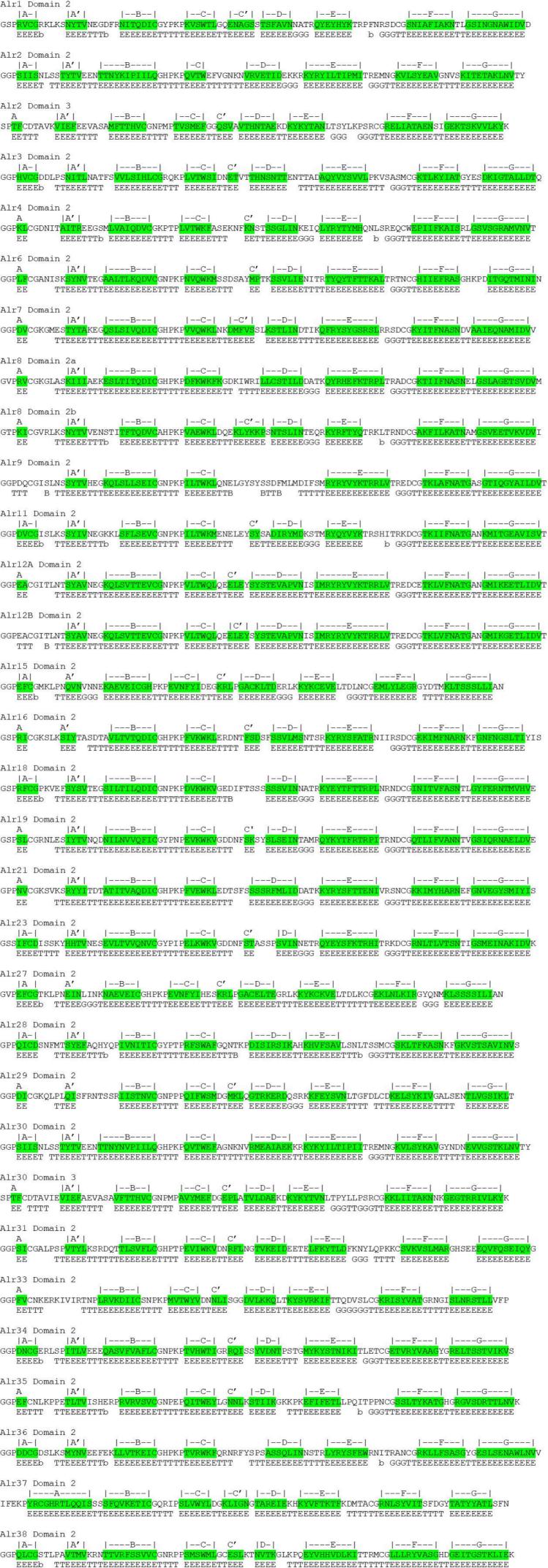
	Conserved V-set residue ensu Cannon et al (2002)	Gly <sup>16</sup>	Cys <sup>23</sup>		Trp <sup>41</sup>			,	Arg <sup>75</sup>	Le (or other h	eu <sup>89</sup> ydrophobic)		Cys <sup>104</sup> r <sup>102</sup>			
		10	20	30 40	50	60	70	80	90 100	110		120	130 14	0 150		160
	CLM2_MOUSE/19-124 CLM5_MOUSE/23-126	TGPE - EVS - GQEQG	SLTVQCRYSSYV	(К <mark>С</mark> ҮМК VК <mark>С</mark> ҮКК	YWCRGVPQRS	CD   LVET [	ОК	SEQLVKK-	NRVSIRDN	QRD-FIFT	VTMED-LR	RMSDAG - I	YWCGITKG	- GPD	- PMFKVN	
	FCAMR_HUMAN/65-171	KGSR-LVS-GEPGG	AVTIQCHYAPSS	SVN RHQRK	YWCRLGPPRWI	CQT I VST I	NQ	YTHHRYR - I	DRVALTDF	PQR - GLFV	VRLSQ-LS	SPDDIG-C	YLCGIGSE	NNM	- LFLSMN	NL - TIS
	FAIM3_HUMAN/19-123 Q8BNV8_MOUSE/24-121	QNPE-LLR-TOEGE	TVSVTCWYDSLY		IWCKQIDNLC			YPFVSKSAEK-	PRELIQOS	SRF - NFFT	VTMTK-LK	MSDSG - I	YHCGIVAN	NTSV	-YLRNIH	1L V \
	TREM1_MOUSE/25-134 CD7_HUMAN/31-132	EEERYD-LVEGQ		YANSQK	AWQRLPDGKE			PFTRPSEVHM-	GKFTLKHD	PSE - AMLQ			YRCVIYH <mark>P</mark>	<b>P</b> ND <b>P</b> VV		
	CD7_HUMAN/31-132 F6RFK3 CHICK/32-139	QTPG-YIL-TKTNN	ISTEIVCPMKG	EHT <mark>GV</mark>	YWYRWNQGRQH	FEFLLFS	8 <mark>P</mark> - L <mark>G</mark> KATY	GTNISQ	E <mark>KFSI</mark> RGT	SSY-HSYRI	LHINR-LH	IGSDNG-T	YYCCTIQS	SQLI	- LGTGTC	2L - DV -
	A0A096MJE4_RAT/42-151	- OPS - VVL - ASSHG	VASEPCEYASSE	1NT DEVRV TSTR I	TVL ROTNDO \	TEVCATTETV	(		NTLGFLDDPFCS	GTENE - SRVN		RAADTG - L	YECKVELMYPPPYE	VG	- MGNGTC	) I Y \
	CD8B_FELCA/24-135 CD8B_RAT/24-134	QTPS-SLL-VQTNQ	TAKMSCEAKTFF	<mark>2</mark> K <mark>G</mark> TT I	YWLRELQDSNKN-	- KH - FEFLASR	Г <mark>Б - ТК</mark> Б   КҮ	<mark>G</mark> ER <mark>V</mark> K-I	KNMTLSFNS	TL <mark>P</mark> F	LKIMD-VK	(PEDSG-F	YFCAMVGS	<mark>P</mark> MVV	- FGTGTK	( L - T V \
	CD8A_MOUSE/36-148															
	B8ZZZ4_HUMAN/29-130 TVA2_MOUSE/26-132	IF PK - KMD - AELGO LDRT WNLGE QSPE - SL I - VPEGA QPDA - RVT - VSEGA TQES - ALT - ISPGE QOPG - AEL - VKPGA HOPP - MKS - SALGT QSPR - YL I - LGRAN QTPR - YL I - LGRAN QTPR - YL - KGGGQ QSPR - HEV - IEMGQ QSPR - SKV - AVTGG	RTSLNCTFSDS -	ASQYF	WWYRQHSGKA	- AS - PIFLLYLS	5 QNK <mark>P</mark> KA	SNGEKEE-	GRFTIHLN		LILSD-FR LHIRD- <mark>S</mark> Q	QPSDSA-L	YLCAVTLY	GGSGNKLI	- FGTGTL	
	TVA1_MOUSE/25-130 LV1B_MOUSE/24-128	QPDA - RVT - VSEGA	SLQLRCKYSYS-	····AT <mark>P</mark> YL	FWYVQYPRQG	LQLLLKY	Y S <mark>G</mark> D <mark>P</mark> VV	QGVNGFE-	AEFS	KSN-SSFH	RKAS-VH	HWSDSA-V	YFCAVSGF	ASALT		( V - 1 V L
	HVM07_MOUSE/24-138	QQPG-AEL-VKPGA	SVKLSCKASGYT	FTSYWM	HWVKQR <mark>PG</mark> RG	LEWIGRIE	PNSGGTKY	NEKFK-	SKATLTVD	KPS-STAY	MQLSS-LT	SEDSA-V	YYCARYDY	Y <mark>G</mark> SSYFDY	- WGQGT	
	VPREB_HUMAN/24-134 TVB8_MOUSE/25-130	HQPP - AMS - SALGT		IDIGVYSV	YWYQQR <mark>PG</mark> H <mark>P</mark>		S-QSDKSQ		PRFSGSKD	VAR-NRGY			YYCAMGAR	SSEK	- EERERE	
	TVB1_MOUSE/25-132	QTPR-YLV-KGQGQ	KAKMRCIPEK	GH <mark>P</mark> VV	FWYQQNKNNE	FKFLINF(	Q NQEVLQ	QIDMTE-I	K <mark>R</mark> F <mark>S</mark> AE	- CPSN-SPCS	LEIQS-SE	AGDSA- L	YLCASSLF	GTSDYT	- FGSGTF	
	TVB1_HUMAN/25-135 TVB5_MOUSE/14-122	QSPR - HEV - TEMGQ	EVTLRCKPIS			LELLIYF	N NNV <mark>P</mark>   D	DSGMPE-I	D <mark>RFS</mark> AK	- MPNA - SFST	LKIQ <mark>P</mark> - <mark>S</mark> E	EPRDSA - V	YFCASSFS	TCSANYGYT -		
	1100 1100000000000000000000000000000000															
	SHPS1_BOVIN/38-145 SIRPD HUMAN/34-140	QPER-SVS-VAAGE OTEM-SOT-VSTGE	TATLHCTVTSLS	PVGP	KWFKGTGPG	REFIYSQI	<pre>&lt; EA<mark>P</mark>F </pre>		PRVTNVSDAT	- KRNN - MDFS	IRISN-IT	PADAG - V	YYCVKFRK	EERGDMEFK-	- SGPGTH	HL-TVS
	HHI 42 HI MAN/226-330	TMKD- GLH- KMOSE	HVSLSCOPVNDY			FSVI AVVI	S-SSONTLINE		SPESWNKEL						EVTI	LTIF
	IGS11 HI MAN/27-141	FYPA-WLT-VSEGA ESPG-SIO-VARGO	PAVLPCTETTSA	SEDLML AL <u>IN</u> LNV	NWNRLSPSNQ	TEKQAAF(	C NGLSQPVQD C GGQMFD	GAPREH-	ARFQIIQL GRVGFTGT	PNR-HDFH	MNILD-TR	RRNDSG - I	Y L CGAISL	- HPKAKIE	- ESPGAE	
Canonical	CXAR_DANRE/28-142	TGQT - SIE - KASGE	SVKLDCQFTLAS	6 D D S <mark>G P</mark> L D I	E <mark>WS</mark> LQ <mark>P</mark> SDNQKE -	EKV <mark>V I</mark> VYS	8 <mark>G</mark> DRAFE	HYYDPLK-	GRVHFNSPD	PKN-GDAS	MNIMG-LK	(ATDTG-T	YQCKIKKV	<mark>PG</mark>	- ASRKYL	L - T V N
	F1P992_CANLF/18-131 VSIG1_CHICK/26-136			2ED Q <mark>GP</mark> L D I 2 <mark>P</mark> L GNFF I												
	JAML_HUMAN/25-136	VSPP-ELT-VHVGD	SALMGCVFQSTE	EDK C   F K	D <mark>WT</mark> LS <mark>PG</mark> EHAKD-	EYV <mark>LY</mark> YYS	8 NLSVPI	GRFQ-I	NRVHLMGD	ILC - NDGSI	LLLQD-VQ	EADQG - T	YICEIRLK	GESQV	- FKKAVV	/ L - H V L
uomains	Q9UEL6_HUMAN/15-129 L8YBJ5 TUPCH/164-276	- TDR - EVH - GAVGS	RVTLHCSFWSSE	6 T T <mark>G</mark> <mark>G</mark> L T S V EWV S D D I S F	TWRYOPEGGRD	A I S I F HY/	4 K <mark>G</mark> Q <mark>P</mark> YI	DEVGAFK-I	ERIQWVGD	PRW- KDGS	IVIHN-LD	SDNG - T	FTCDVKNP	PDIVG	- KTSQVT	L Y\
	SCN3B_MOUSE/28-143	EV <mark>P</mark> S-ETE-AVQGN	ISMKLRC I <mark>S</mark> CMKF	REEV EATTVV	E <mark>WF</mark> YR <mark>P</mark> EGGK	DFL IYEYI	R N <mark>G</mark> HQEV	ES <mark>P</mark> FQ-0	G <mark>R</mark> LQWNGS	KDL-QDVS	ITVLN-VT	IN <mark>DSG</mark> - L	YTCNVSRE	FEFEAHR <mark>P</mark>	- F V K T <mark>T</mark> F	RL - I <mark>P</mark> L
	SCN1B_HUMAN/23-139 A2BGB5 DANRE/30-142	- TSS-ELE-AVYGM	ITFKILCI <mark>S</mark> CKRF EIRLKCTFKSNF	RSET NAET F <mark>T</mark> R <mark>P</mark> LS EERT SV	EWTFRQKGTEE SWSFMPLGKTK	FVKILRYE EE <mark>PFF</mark> HY(	E NEVLQL Q <mark>G</mark> HAYL	EEDERFE-	GRVVWNGSRG DHVVWSGD	- TKDL - QDLS VMK - GDGS	IFITN-VT ITLQD-VQ	TYNH <mark>SG</mark> -D DFSFNG-T	YECHVYRLLFF YSCQVLN <mark>P</mark>	PDIQG	- ENYEHN - FAGEIK	NT - SVN KL - RV -
	MPZL2_HUMAN/30-143	EVDS-ELE-AVYGM -TSS-ELE-AVNGT YTSR-VLE-AVNGT TVPA-TLN-VLNGS KVPTEPLS-TPLGK QVQK-SVT-VQEGL	DARLKCTFSSFA	PVG DALTV	TWNFRPLDGGP	EQFVFY	1ID <mark>P</mark> FQ	<mark>P</mark> MS <mark>G</mark> R <mark>F</mark> K - I	DRVSWDGN	PER - YDAS	ILLWK - LQ	FDDNG - T	YTCQVKN <mark>P</mark>	PDVD <mark>G</mark>	- VIGEIF	RL - SVV
	SCN2B_HUMAN/33-147 VSIG2_HUMAN/27-142	KVPTEPLS-TPLGK	TAELTCTYSTSV	/GDSFAL	EWSFVQPGKPI	- SE - SH <mark>P</mark> ILYF1	R ΜΚΙΙΝL Γ Ν <mark>G</mark> HLY <mark>P</mark>		KRVSLLQNP	PSK-YDVS	VMLRN-VQ LKLTD-VH	PEDEG - I PSDTG - T	YLCQVNNP	PDRHR	- GLGLIN	
	SIGL5_HUMAN/24-140	QVQK - SVT - VQEGL	CVLVPCSFSYPV	VRSWYSS <mark>PP</mark> LYV	YWFRDGEIPY	YAEVVAT	N NPDRRV	K <mark>P</mark> ETQ-	GRFRLLGD	VQK - KNCSI	LSIGD-AR	RMEDTG - S	YFFRVER <mark>G</mark>	RDVKYSY	- QQNKLN	NL - EVT
	TIMD4_HUMAN/23-128 Q9R121_RAT/32-135	IELV - PPM - VAEGG	NSVLFVHEMP	SHNSNS <mark>M</mark> LNVQAF	YWYKQRD <mark>P</mark> TK	SYEVARYI	_ T - PTNESS	KMPQHS-	GRKTVF	YSGSI	LLIRN-VT	rqa <mark>dsg</mark> - v	YTLLTFNTEMQS		- ELTHVH	IL-EVF
	Q3LFS9_MOUSE/42-141 Q9JHL7 RAT/42-140				AMAYICANTTAL	DKELADEL	D NICHMANE	TOOANE	CRELLY	CNCC	L FOM LT				TOAT	
	CEAM4_HUMAN/38-141	IEAL - PSS - AAEGK	DVLLLACNIS	ETIQAY	YWHKGKTAEG	SPLIAGY	IT - DIQANI	PGAAYS-	GRETVY	PNGS	LLFQN-IT	LEDAG-S	YTLRTINASYDS		- DQATGC	AL-HAH
	CEA21_HUMAN/38-141 M0QX98_HUMAN/6-107	IASA - PFE - VAEGE		QEFQVF ETIQAY QHLYSY	GWYKGKTVEP	NQLIAAY	/I-DTHVRT	PGPAYS-	GRETIS		LHFQN-VT		YTLQVTYRNSQI		- EQASH	
	CD2_RAT/23-120	RDSG - TVW-GALGH	IGINLNI PNFQMT	「 DD I DE 🛛	RWER <mark>G</mark>	STLVAEF	(RKMK <mark>P</mark> F	LKSGAFE-	L	A NGDI	LKIKN-LT	rrddsg - t	YNVTVYST	NGTRI	- LNKALC	
	CD2_MOUSE/23-121 CD2_HORSE/25-122	RDNE - TIW- GVLGH		「DDIDE <mark>V</mark> SEHVED <mark>I</mark>	RWVRR	GTLVAEF	(RKK <mark>PP</mark> F	LISETYE-	VL	AN <mark>GS</mark> I		IRN <mark>DSG</mark> - T	YNVMVYGT	NGMTR	- LEKDLC	
	CD2_HUMAN/28-127	TNAL - FTW- GALGO		S DD   DD	KWEKTSD	KKK I AOFE	R KEKETE	KEKDTYK-	I F	KNGTI			KVSIYDT	KGKNV	- I EKIED	) - K I C
	F1NMT9_CHICK/44-146 SKIT3 MOUSE/32-141	RVT-ANVGQ		(DA WRLDI	RWIQQRS	SGFVHHY	Q DGEDLE	QMEEYK-	GRTELLRDG		LRITA - VS	SSDSG - S	YSCAVQDGDAY			NL - EV -
	BTNLA_MOUSE/37-146	GPPH-PLL-AIVGQ	DKELPCKLSLNI	SAEGMEL	RWYRDK <mark>P</mark>	SSVVHVY	(NGEDVY	DEQMVEYK-	GRTSFNGSH	VAR - GEAA	VKIHN-VT	FVFDNG-T	HCVFKEYTSH	· · <u>-</u> · <u>-</u> · · · · · · · · ·	- SQATLY	VL - KVA
	NECT4_HUMAN/35-146 MO2R1 MOUSE/42-147	GLER-PVL-APLGG GPPH-PLL-AIVGQ ETSD-VVT-VVLGQ QVNT-TVS-VQIGT NST-GVLGG	DAKLPCFYRGDS	S <mark>G</mark> E QV <mark>G</mark> QV	AWARVDAGEG	AQELALLI	HS - KY <mark>G</mark> LHV / DTKTNE	SPAYE-	GRVEQPPPP		VLLRN-AV	QADEG-E		<mark>P</mark> AGS	- FQARLF	
	Q8K094_MOUSE/37-141		STTLHCSLTSNE	NV T I TQ I	TWMKKDSGGS	HALVAVF	H <mark>P</mark> - KK <mark>GP</mark> NIKEP		ERVKFLAAQ	QDL - RNASI	LAISN-LS	SVEDEG-I	YECQIATE	<mark>P</mark> R <mark>G</mark>	- SRSTNA	W-LKV
	CSPG2_MOUSE/27-148 GPA33_HUMAN/25-138	ETSP-PVK-GSLSG ETPODVLR-ASOGK	KVVLPCHFSTLF SVTLPCTYHTST	TL <mark>PP</mark> NYNTSEFLRI SSRE <mark>G</mark> LI	KWSKMEVDKNGKE OWDKLLLTHTE	DIK-ETTVLVAC	2 N <mark>G</mark> NIKI 5 NKNYIH	GQDYK-	GRVSVPTH	PDDVGDASI	LTMVK - LR	RASDAA - V MADNG - T	YRCDVMYGIED		- TQDTMS	
	MXRA8_HUMAN/168-292	DGEK-EVLAVARGA	PALLTCVNRGHV	/WTDR-HVEEAQQVV	HWDROPPGVPHDF	RA DRLLDLY	A SGERRA	YGPLFLR	DRVAVGADA	FER-GDFS		EVADEG - T	YSCHLHHHYCGLH-		- ERRVFH	IL - TVA
	Alr01.1_D1/5-115 Alr02.1_D1/4-114	TSPA-ILE-EVVGR	SVTITYVTDVAD	SEATASM DVADDVVVNFRT	N <mark>Y</mark> NDSRIAE <mark>G</mark>	L I T I L T <mark>G</mark> / T K T <mark>V F</mark> D	a V D <mark>G</mark> FNV	AQGGRKLFG - I KI <mark>PP</mark> TPFG - I	NRLR <mark>T</mark> STPL		LILQN-IQ LLIDN-LE	EYNETL-T	FULLVGSSTLV FFAKIEIF	K <mark>P</mark> NVE	- KAANIS - ANSTIN	5 I V E I S N L - I I Y
	Alr03.1_D1/5-119	TDSP-TIS-GIFNQ	SAIISFKATITA	NDNRTVKSF	YVEL <mark>PP</mark> S		S NNVLQA	LPVPPFN-	SRLTASAAN	QEYL	LQVN <mark>P</mark> - LK	FSDDY-T	FRGVLIYL	LNNDINNPI-	AVEQDVK	(L - DVF
	Alr04.1_D1/6-120 Alr06.1_D1/5-115	TDSP-TIS-GIFNQ NGQS-EVT-GMLDK AVKT-NFE-VPVNQ	TAQLQWRVVPDA	AGEMLQSH	EVFVLGLPN	VQ11		TIAGKERFG-I	NRISGSLTN	ADA - AVF KI	LSIKK-IQ	QFNEKK-S	FNLKVVFY	KE <mark>P</mark> DYY <mark>P</mark>	- KNSTVV	
	Alr07.1_D1/8-125 Alr08.1_D1a/7-122	VVKN-VFERVQIKS	TVVMEWKIAHSD	0 NQTNKVL 2 AEQIVVL	K <mark>LY</mark> VL <mark>P</mark> DRD	۰ R <mark>PVF</mark> SS۱	( <mark>G</mark> KYQSS	QGKGRQTFG - I	N <mark>RLSA</mark> TFSKIK -	GKYT	VTLKR-IQ	YNENY - T	FQLKVIFI	NKKSVTET	- KVADIF	RIKNVN
	Alr08.1_D1b/8-119	DTAS-TLD-IVYNT	TLKLDWNMTLSA	λ <mark>G</mark> EV I V <mark>GV</mark>	Q <mark>VY</mark> VL <mark>P</mark> DTT	NR I I TD1	Г -  -  -  -  S <mark>Р</mark> Т V -  -  -	LSKGISIFGEI	N <mark>RLS</mark> ATFTN	SRYRI	LMLRN-IR	RYNESF-T	FQLYVIYG	AGDHLSS	- SRFNIC	2 V - T V H
	Alr09.1_D1/5-114 Alr11.1_D1/5-114	AIIK-NIE-ATDNS	TAELSWRVETNK	(D <mark>G</mark> ERVF <mark>GV</mark> / <mark>G</mark> EE I STA		· · · · · · · · · · · · · · · · · · ·	D DKS <mark>G</mark> TV	TEAGRNKFG-	GRLSASFSN	NVYK			FTLTAAFY	WKSALDP	- VNDTAT	ITSV
	Alr12A.1_D1/1-110			A GEOVEGV	EVEVI GSTN	VOII	KI TSTI	TISGKKREG-I	NRISGSIRN	GIVTI			ENIKVVEY	KAPEYYP	- KNSTVV	
	Alr12B.1_D1/1-112 Alr15.1 D1/8-121	- VKT - NLE - VPVNQ		GEKITTV	NVFILES <mark>P</mark> K	VL <mark>IV</mark> L <mark>G</mark>	[HLASAT	TPAGKTMFG-I	DRLSGSLSN		LSIKK-IQ	OFSEEK-S	FNLEVLFE	SQP-FYL	- KNATVV	
Alr	Alr15.1_D1/5-118	ISTS - FLE - ATYGN	IT V DMHWK I I L DT	GQTINSF	TL <mark>S</mark> LKSR <mark>P</mark> E	D <mark>S I I</mark> F <mark>G</mark> S	6 ANYQNI	ANKGKELFG-	NRLSAVYDKTT-	<u>S</u> TYR	VSLKN-IQ	YNETL-S	FQLT <mark>T</mark> TLS	NPLFG	- KQTTIE	IKDV
	Alr17.1_D1/7-120 Alr18.1 D1/5-116	PVKS-VLE-VTHGS PVLP-TIS-TTENA		GEKITTV GQTINSF 	SLVVLPDVG TVRELPS	· · · · · NPVVSGI	N VN SQMV F VD <mark>G</mark> ENV	QEKGRELFG-I	NRLSATENKTA - DRISGIENNSS -	GLYI	ATLGN-IQ		FRLM <mark>T</mark> TFS FELVVVSSKFVI	L <mark>P</mark> DQL	- EGNIIE	
	AIF19.1_D1/5-115	PVLP-IIS-LIFNA			IVRELPS	L I I <b>I L</b> I 🖸 /	4 V DG F N V	VQGGRKLFG-I	DRVSGIFNNRN-	NIVI			FQLLVGSSILV		- KAANIS	DIVEIS
	Alr21.1_D1/5-118 Alr23.1_D1/7-115	PIAM-TKC-ATVNT TKSN-NKI-VLNGN	TVKLTWKVKLE	2 SE <mark>G</mark>     Y <mark>  </mark> 2 <mark>G</mark> D     A L R	KLHKLPDDE SVYLSPDFI		A SQKLTV	IEKGRELEGG	KRLSANYSKHY - NRLSATYGN	ENVTI	LTLQN-IK MSLMD-VR		FYFVARFF	PGFVL	- GTGAIT	
	Alr27.1_D1/6-119	DANN - KNE - ASKGK	DY IVHWNISNAE	KIFL	E <mark>L S</mark> K N K T N I K		SNN <mark>PG</mark> N	WKSLELFTV	NKTSISIETFM-	- EGDK - MCCI	MTVHN-VS	SYDDDASE	YVMQVTYI	LYNPIREEK-	- KNATYF	RL - DVT
	Alr28.1_D1/7-118 Alr29.1 D1/7-111			(RTYKN E												
	Alr30.1_D1/3-109	ISPA-ILE-KVVGR	SVT I T YV <mark>T</mark> DVAD	DNIDAIFK	Y <mark>Y</mark> NDSQ   <mark>G</mark> E <mark>G</mark>	TKRLFE-		TPPTPFG-Y	VRLRTSTSL	QNQ - KKYTI	LHIDN-LK	KYNDTG - L	FFAKIEIF	K <mark>P</mark> NVE	- ANSTTY	(L - 1 1)
	Alr31.1_D1/3-110 Alr33.1_D1/10-118			/KNQR <mark>L</mark>												
	Alr34.1_D1/5-118	I I KS - QIT - GVVRE	SVDISFIIDTEN	N <mark>G</mark> ESLLS <mark>V</mark>	K <mark>VF</mark> QL <mark>P</mark> DK	LN <mark>PVA</mark> V <mark>G</mark> S	6 SSAFTV	AKGGSFG-	GRIQANVE	PAK - GRYTI	LAIST-LK	KYVDEA-A	YEITATFY	ESTAEVRV	- LRKT IC	DL-KVC
	Alr35.1_D1/5-119 Alr36.1_D1/5-116	AISN-SMT-AKLNS	TVVLQWKLFIF	2 EEKFNVI RMDEI	RLYAEPNLVEP	L I K <mark>V</mark> RYYF	R R <mark>G</mark> VVEI	LNASIEMFG-I	NRISATYNG	SLFT	VLLNK-VV	/YTDSH-K	YLLRVMKYYFPMR-		- LQSTVT	<sup>-</sup> L - NVT
	Alr38.1_D1/10-122	VVQQ - DL I - GLKGE	DVEVTLITTMLV	/GQRLLS	KVFQL <mark>PG</mark>	NIELATG	6 ERAFLL	QNETLR <mark>G</mark> -	IKFETKID	GEK - GIYRI	FGIKS-LD	DT - M	YLATAIFH	D <mark>G</mark> YTRYHK	- DDVR <mark>V</mark> K	(V - NV L
Co	onserved V-set residue								1							
	nsu Cannon et al (2002)	Gly <sup>16</sup>	5 Cys <sup>23</sup>		Trp <sup>41</sup>			,	۲ Arg <sup>75</sup>	Le	u <sup>89</sup>	Asp <sup>98</sup> Ty	r <sup>102</sup>			
		Siy			-				-	(or other hy			Cys <sup>104</sup>			
													0,0			

**Fig. S12. Multiple sequence alignment of V-set Ig domains and Alr domain 1.** Alr domain 1 sequences V-set Ig domains from pfam (pf07686). The positions of conserved V-set residues according to the nomenclature of Cannon et al [30] are shown above and below the alignment. Residues are highlighted by sequence conservation and chemical property with CLUSTALX colors as implemented in Jalview.



# Fig. S13. Measurements of the turn from strands A' to B in the predicted structures of domain 1.

Left graph shows distance between the alpha carbons of the first and fourth residues in the turn. In a beta-turn, the distance between these two atoms is less than 7.0 Å. Turns <7.0 Å are shown in black bars. Those >7.0Å are showing with gray. Right graph shows distance between the oxygen atom of the first residue and the nitrogen atom of the fourth residue. A distance of <3.5Å is considered small enough for hydrogen bonding to occur. Black bars indicate distances <3.5Å. Gray bars indicate distances >3.5Å. Gene names are color-coded according ot whether they are *bona fide* (blue) or putative (orange) genes.



# Fig. S14. STRIDE secondary structure predictions for Alr domains 2 and 3

For each domain, the top line shows beta-strands labeled according to their position in the primary amino acid sequence. The middle line shows the sequence of the domain. The bottom line shows the STRIDE secondary structure predicted from the Colabfold model. (H = alpha helix, G = 3-10 helix, I = PI-helix, E = beta-strand extended conformation, B = isolated bridge, T = turn.) Figure S9 – I-set alignment

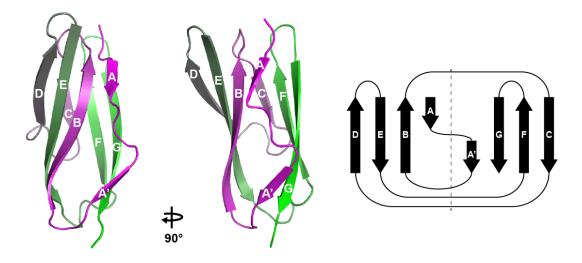


Fig S15. Predicted structure of Alr2 domain 2.  $\beta$ -strands labeled on the predicted structure of Alr2 domain 2. The corresponding Greek key is shown to the right.

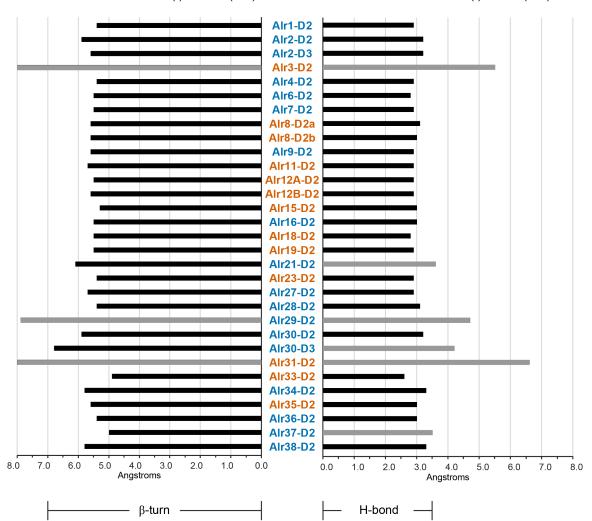
	I-set motifs (Wang 2013)	ix. Pro 	iii. ii. Gly Cys I I	v. iv. cis-Pro Trp 		viii. Pho-X-Pho-XX-Pho	i. i. vi. vii. Tyr Asn Pho-X-Pho-X  Cys
	V-frame residues (sensu Cannon <i>et al.</i> 2002	2)	Gly <sup>16</sup> Cys <sup>23</sup>	Trp <sup>35</sup>	Arg <sup>75</sup> (or Lys/His)	Leu <sup>89</sup> Asp <sup>98</sup>	Tyr <sup>102</sup> Cys <sup>104</sup>
	CAVPT_BRALA/59-149 HMCN1_HUMAN/1171-1258	<mark>P</mark> KIQR <mark>GP</mark> KH L KV	Q V <mark>G</mark> Q R V D I <mark>P</mark> C N A Q <mark>G</mark>	TPL - PVITWSK <mark>GG</mark> S	<u>T</u> - <mark>M</mark> L - VD <mark>G</mark> E <mark>H</mark> H <mark>V</mark> <u>-</u> SN <mark>F</mark>	<mark>?</mark> - D <mark>G T L S I DQ - A T P S D</mark> A G	90 100 91 - YTCQATN - SFGEA - FDSARLAL - 31 - YTCVATN - IAGTD - ETEITLHV -
	HMCN2_HUMAN/4303-4389 IGDC3_HUMAN/330-417 LRIG1_MOUSE/601-691 MUSK_HUMAN/121-208	AEFVQH <mark>P</mark> QS ISR PSFTKIPHD IAI PKITRPPIN VKI	<mark>P</mark> AGTTAMFTCQAQG RTGTTARLECAATG I <mark>E</mark> GLKAVL <mark>P</mark> CT <mark>T</mark> MG	BE <mark>PP - PHVTWLKNG</mark> G HPN - PQIAWQKDGG NPK - PSVSWIKGDS	V - LG - <mark>PGG</mark> H VRL KNI TDF	N - N <mark>ST</mark> LTIS <mark>G</mark> - IGPEDEA D - DDVFFITD - VKIDDMG <mark>S</mark> LRIHN - VQKEDAG	T - YDCVAHN - LLGSA - TARAFLVV - AI - YQCVAEN - SAGS <mark>S</mark> - QASARLTV - V - YSCTAQN - SAGSV - SANATLTV - Q - YRCVAKN - SL <mark>G</mark> TAYSKVVKLEV -
	MYLK_CHICK/28-118 MYLK_CHICK/156-245 MYLK_SHEEP/331-421 MYOM1_MOUSE/1555-1645	<mark>PKFA</mark> TK <mark>P</mark> NR VVV PYFSKTIRD LEV	R <mark>EG</mark> QT <mark>GRFSCKITG</mark> V <mark>EG</mark> SAARFDCKIEG	RPQ - PQVTWTKGDI YPD - PEVVWFKDDG	H - LQ - QNER FNM FEKT S - IR - ESRH FQI DYQEI	- IQYLEIQN - VQLADAG OGNCSLIISD - VC <mark>G</mark> DDDA	SK - YTCEAAN - D <mark>GG</mark> VR - QVTVELTV - SI - YTCTVVN - SAGKA - SVSAELTV - AK - YTCKAVN - SLGEA - TCTAELIV - SK - Y <mark>G</mark> LVVKN - KYGS <u>E</u> - TSDFTVSV -
	NCAM2_HUMAN/21-111 000557_HUMAN/20-114 055124_MOUSE/1345-1435 057576_CYNPY/213-301	LQVDIV <mark>P</mark> SQGEI KVI <mark>GG</mark> L <mark>P</mark> DV VTI	SV <mark>G</mark> ESKFFLCQVAG M <mark>EG</mark> KTLNLTCTVFG	OAKDKDISWFS <mark>P</mark> NG NPD- <mark>PEV</mark> VWFKNDK	EKLT - <mark>P</mark> NQQR I <mark>S</mark> VV WNDDS D - I E - L SEH FLV KMEQS	S - S <mark>ST</mark> LTIYN - ANID <mark>D</mark> AG SKYV <mark>S</mark> LTIQ <mark>G</mark> - VTAEDSG	I - YRCQATD - AKGQ <b>T</b> QEATVVLEI - I - YKCVVT <mark>G</mark> - EDGSESEATVNVKI - K - YSINVKN - KYGGE - KIDVTVSV - D - YTCIAEN - KAGEQ - EASILLKV -
	076281_DROME/6357-6446	PEF <mark>T</mark> KPLHDLTI PHIDRVNLK-PVIV	H <mark>DG</mark> EQLILTCYVK <mark>G</mark> KT <mark>G</mark> LSISLDINIRG	BDPE - <mark>PQISWSKNG</mark> K BEPA - <mark>PKV</mark> EWFFNNS	S - L S - SSD I L DL RYKNC S - VT - SDEH - SVKI DNVD	- I A <mark>T</mark> L T I NE - V F <mark>P</mark> E D E G ′ - N <mark>T</mark> K F F VMR - AQ R SQ SG	K - YTVTAAN - EFGKD - TADIEVIV - V - ITCTATN - SVGAV - ETKCKLTI - K - YIIKATN - EVGED - EAELEVTV - Q - YVCRARN - AIGEA - FAAVGLQV -
Canonical	P79757_CHICK/328-417 P79757_CHICK/734-823 PALLD_HUMAN/1135-1225 PALLD_HUMAN/271-361	<mark>P</mark> AFITQ <mark>P</mark> KS QNV <mark>P</mark> HFLQA <mark>PG</mark> D LTV	N <mark>EG</mark> QDVLFTCEVSG Q <mark>EG</mark> KLCRMDCKVSG	DPS- <mark>PEVEWLRNNG</mark> LPT-PDLSWQLD <mark>G</mark> K	<mark>P - I A - VSSH MRA TRSKI P - VR - <mark>P</mark>D<u>S</u>A HKML - VREN<mark>C</mark></mark>	Ŋ - T <mark>YS</mark> LEIRN - AAVSDTG S - VHSLIIE <mark>P</mark> - VTSRDAG	S - YTVVVEN - SEGRQ - EAHFTLTV - SK - YTVKAKN - YHGQC - SATASLTV - GI - YTCIATN - RAGQN - SFSLELVV - SR - YTCLATN - PSGSD - TTSAEVFI -
l-set domains	PXDN_HUMAN/433-519 Q08476_CHICK/372-461 Q08476_CHICK/180-269 Q8BID4_MOUSE/6-97	<mark>PFFDTPITP VDG</mark> PYFVTPLEP VQV	II <mark>G</mark> ESADFECHISG TV <mark>G</mark> DSASLQCQVAG	TQ <mark>P</mark> - IRVTWAKDNG TPE - MIVSWYK <mark>G</mark> DT	E - I R - T <mark>GG</mark> N <mark>Y</mark> Q I SYVE1 K - L R - <mark>G</mark> TAT VKM HFKNO	N - TAHLTILR - VDR <mark>GD</mark> SG 2 - VA <mark>T</mark> LVFSQ - VDSDDSG	Q - YECQAVN - IIGSQ - KVVAHLTV - K - YTCYASN - EVGKD - SCTAQLNV - E - YICKVEN - TVGEA - TSS <mark>S</mark> LLTV - R - YSLRATN - <mark>GSG</mark> QA - TSTAELLV -
	Q8T103_BOMMO/4625-4714 Q90X22_DANRE/178-268 Q90X22_DANRE/66-155 Q90YM0_DANRE/209-296	<mark>PCFT</mark> TQIQ <mark>P</mark> VQC PVFESKLTP AEV VSVPQQ <mark>S</mark> FN ATA	V <mark>EG</mark> SEVKFQ <mark>Y</mark> KVTG TIGESVRFTVTVSG D <u>Y</u> GESVTFTCRAYG	TPF-PDVQWFKGNS FPK-PKVQWFHNGK SPE-PDVTWHRKGV	Q - I K - SSQ T <mark>C</mark> SV VCN <mark>P</mark> I A - I T - SSS I <mark>Y</mark> T F VEERI Q - L Q - E <u>S</u> ER <mark>Y</mark> VM RAR <mark>G</mark>	D <mark>G</mark> SGFLIMSN-IQQRDSG D-E <mark>YS</mark> LIITK-VKKDYEG TTLTVRN-IQQDDGG	BD - YKCVAYN - SAGRV - TVAAKLKV - BL - YTCKAVN - <mark>P</mark> FGEA - SCSAELIV - BE - YSCTASN - RFGQ <mark>T</mark> - TCK <mark>T</mark> ILKV - BS - YTCRASN - KAG <mark>EV - EHELFLKV -</mark>
	Q90Z41_CHICK/223-308 Q95YM1_PROCL/4632-4721 Q98918_CHICK/3072-3160 Q9GV22_MYTGA/344-433	<mark>P</mark> KIVQKLKS QVV PAVIV <mark>P</mark> LRD AV <mark>T</mark> PRF <u>I</u> RK <mark>P</mark> RN ILA	Q <mark>EG QG</mark> A F F E C I I T A S EG QS A R F Q C R V T G A EG QS <mark>T</mark> K F D C K I I G	S <mark>P</mark> K - <mark>PK I QWMKG</mark> KA 3T - D - L KVSWYSKDF 3A <mark>SP</mark> - <mark>P</mark> I VTWSYDNS	<mark>P</mark> - I K - QSKY F TM TAD <mark>G</mark> I E - I K - <mark>P</mark> SRF F RM TQFEI V - L S - QSVK <mark>Y</mark> MQ KYR <mark>G</mark> I	D - R <mark>YT</mark> L R I SE - AF <mark>P</mark> EDEG D - TYQL E I AE - AY <mark>P</mark> EDEG N - EY <u>E</u> L K I SR - L KMADKG	T - YTCIAEN - RVGKV - EASATLTV - T - YFCVATN - <mark>PS</mark> GKC - TVEAKLQV - T - YTFVASN - SVGQV - TSTAILKL - SL - YTVIAEN - <mark>SFG</mark> KR - EEH <mark>A</mark> TLKV -
	Q9GV22_MYTGA/647-736 ROBO1_RAT/262-347 SPEG_MOUSE/727-816 SPEG_MOUSE/1490-1579	<mark>P</mark> SFVKR <mark>P</mark> SN LAV PVFE I PLQN MVV PRFES I MED VEV	T V D D S A E F K C E A R G A P G A D V L L K C I I T A G P G E T A R F A V V V E G	DPV - PTFGWRKDDG NPP - PQVSWKKDGS KPL - PDIMWYKDEV	E - L <mark>P</mark> - KSR <mark>Y</mark> E I RI M - LH - SE <mark>G</mark> R LL I RAE <mark>G</mark> L - LA - ESNH VSF VYEEI	D - DHTLKIRK - VTAGDMG E - RHTLLLRE - AQAADAG N - E <mark>CS</mark> LVLLS - A <mark>G</mark> SQDGG	SN - YVCKAKN - <mark>PGG</mark> QA - SSRTSVQV - SS - YTCVAEN - MVGKA - EASATLTV - SS - YTATATN - ELGQA - TCA <mark>S</mark> SLAV - SV - YTCTARN - LAG <mark>EV</mark> - SCKAELSV -
	TITIN_HUMAN/7385-7474 UNC22_CAEEL/4150-4239 UNC22_CAEEL/6863-6953 UNC22_CAEEL/6865-6674	<mark>P</mark> KILTA <mark>S</mark> RK IKI PSFTAQLSD SET PRFIVKPYG TEV	KA <mark>G</mark> FT <mark>H</mark> NLEVDFIG EVGGSAEFSAAVSG GEGQSANFYCRVIA	GAPD - PTATWTVGDS GQPE - PL I EWLHNGE AS <mark>SP - P</mark> VVTWHKDDF	<mark>G</mark> - AA - L A <mark>P</mark> E L L V DAKSS R - I S - ESDS - RF RA SYVA E - L K - <u>Q</u> SVK <mark>Y</mark> MK RYN <mark>G</mark> I	S - T <mark>TS</mark> IFF <mark>P</mark> S - AKRADSG - KATLRISD - AKKSDEG N - DYGLTINR - VK <mark>G</mark> DDKG	E - YHCKATN - EVGSD - TCSCSVKF - N - YKLKVKN - ELGED - EA I FEVIV - Q - YLCRASN - SAGQE - QTRATLTV - E - YTVRAKN - SYGTK - EE I VFLNV -
	UNC89_CAEEL/3920-4010 UNC89_CAEEL/648-737 UNC89_CAEEL/3482-3573 UNC89_CAEEL/3286-3377	<mark>P</mark> AFVTKLRD KEC PEF <mark>T</mark> QKLR <mark>P</mark> LEV <mark>P</mark> TFVRELVT <mark>T</mark> EV	K <mark>EG</mark> DVIDFECEVEG R <mark>E</mark> QETLDLKVTVIG KI <u>N</u> ETATLSVT <u>V</u> KG	WPE - PEL VWL VDDG TPV - PNVEWFKDDK VPD - PSVEWL KDG	<mark>P - L R - P</mark> SHD F RL QYD <mark>G</mark> P - I N - I DNSH I F AK DEG S P - VQ - T <u>D</u> SSHV I AK VEG S	2 - T A K <mark>L E I RD - AQ P D D T G</mark> 3 - HHT L T I KQ - A R G E D V G 3 - S <mark>Y S I T I K D</mark> - A R L E D S G	DE - YRCEASN - EFGDV - WSDVTLTV - V - YTVKIQN - EFGSI - ESKAELFV - V - YTCKATN - EAGEA - KTTANMAV - K - YACRATN - PAGEA - KTEANFAV -
	Alr3-F-1_D2/1-100 Alr4-F-1_D2/1-99	<mark>G S P R V C G</mark> R K L K S N <mark>Y</mark> T V G G P H V C G D D L <mark>P</mark> S N I T L <mark>G G P</mark> K L C G D N I T - A <mark>I</mark> T R	N <mark>EG</mark> DFRNITQDICG NATFSVVLSIHLCG E <mark>EG</mark> SMLVAIQDVCG	SY <mark>P</mark> K - <mark>PKVSWTLG</mark> QE SRQK - PLVTWSIDNE SK <mark>P</mark> T - PLVTWKFASE	N A - <mark>G</mark> SST <mark>S</mark> FAVNNATRO T - <mark>V</mark> TTHNSN TTENTTADAO KN <mark>F</mark> K - <u>N</u> STS <mark>SGL</mark> INKEIQI	}YE <mark>YHYKTRP</mark> - FNRSDCG }YV <mark>YS</mark> VVL <mark>P</mark> K - VSASMCG .YR <mark>YTYMH</mark> QN - LSREQCV	HA - YL CT AT C - <mark>G S</mark> KKL - EK <mark>G</mark> I Q VE I - S NIAFIAK <mark>N</mark> - TL G SI - NG NAWID VD KTL KYIATG - YESDK - I G TALLD T VEPIIFKAIS - RL G SV - SG RAMVNVT SHIIEFRASG - HKPDI - TGQ TMININ
	Alr7-F-1_D2/1-99 Alr8-F-1_D2a/1-100 Alr8-F-1_D2b/1-100	GGPDVCGK <mark>G</mark> MESTYTA GVPRVCGKGLASKIII GTPKICGVRLKSN <mark>Y</mark> TV	K <mark>EG</mark> QSLSIVQDICG AEKESLTITQDICG V <mark>E</mark> NSTITFTQDVCA	HPK - PVVQWKLNKC HPK - PDFKWKFKGC HPK - PVAEWKLDQE	M - FV - SSLK <mark>S</mark> TL I NDT I KO K IWR - I L L C <mark>S</mark> T I L DDATKO KL <mark>Y</mark> K - K <mark>P</mark> SN <mark>T</mark> S <u>L</u> I NTEQRI	ΩFR <mark>YSYGS</mark> RS - L RRSDCG ΩYRHEFKTR <mark>P</mark> - L TRADCG KYRF <mark>TY</mark> QT <mark>R</mark> K - L TRNDCG	KYITFNASN - DVAAI - EQNAMIDVV KTIIFNASN - ELGSL - AGETSVDVM SAKFILKATN - AMGSV - EETVKVDVI SKVLSYEAVG - NVSKI - TETAKLNVT
	Alr9A-F-1_D2/1-101 Alr11-F-1_D2/1-101 Alr12A-F-1_D2/1-101	GGPDQCGI <mark>S</mark> LNSSYTV GGPDVCGISLKSSYIV GGPEACGITLNTSYAV	H <mark>EG</mark> KQLSLL <mark>S</mark> EICG NEGKKLSFLSEVCG N <mark>EG</mark> KQLSVT <mark>T</mark> EVCG	NPK - PILTWKLQNE NPK - PILTWKMENE NPK - PVLTWQLQEE	L <mark>GY</mark> S-YSSDFMLMDIFSMF LEYS-YSADIRYMDKSTMF LE <mark>Y</mark> S-YSTEVA <mark>P</mark> VNISIMF	₹Y RYVYK <mark>T</mark> RRL VT RE DCG ₹YQYVYKT RSH I T RK DCG ₹Y RYVYKT RRL VT RE DCE	TKLAFNATG - ASGTI - QGYAILDVT TKIIFNATG - ANKMI - TGEAVISVT TKLVFNATG - ANGMI - KGETLIDVT TKLVFNATG - ANGMI - KGETLIDVT
Alr D2	Alr15-F-1_D2/1-92 Alr16-F-1_D2/1-100 Alr18-F-1_D2/1-99	GGPEFCGMKLPNQVNV GSPRICGK <mark>S</mark> LKSIYTA GSPRFCGPKVEFSYSV	NNEKA E <mark>V</mark> EIC <mark>G</mark> SDTAVLTVTQDICG TEGSILTILQDICG	HPK - PEVNFYIDEG HPK - PFVKWKLERD NPK - PDVKWKVGED	KRL <mark>P</mark> <mark>G</mark> A <mark>C</mark> KLTDERLKH NTFS - DSFS <mark>S</mark> VLMSNTSRH I - FT - SSSS <mark>S</mark> SVINNATRH	<pre>KYKCEVE L T D L N CG KYRYSFATRN - I I RSDCG KYEYTFTTRP - L NRNDCG</pre>	EMLYLEGRG - YDTMK - LTSSSLIA EKIMFNARN - KFGNF - NGSLTIYIS INITVFASN - TLGYF - ERNTMVHVE QTLIFVANN - TVGSI - QRNAELDVE
	Air20-F-1_D2/1-100 Air21-F-1_D2/1-99 Air23-F-1_D2/1-99 Air28-F-1_D2/1-96	GTPETCGESLKPAYNI GPPNVCGKSVKSRYYI GSSIFCDISSKYHHTV GPPQICDSNFMTSYEF	IETANLTVTQDICG TDTATITVAQDICG NESEVLTVVQNVCG A <mark>Q</mark> HYQ <mark>P</mark> IVN <mark>I</mark> TICG	HPK - PDVKWKLEED         HPK - PFVEWKLEDT         YPI - PELKWKVGDD         YPT - PRFSWAFGQN	STFS-SSFS-S <mark>S</mark> VLINNA <mark>P</mark> R S-FS-SSSR-FMLIDDATK N-FS-TASS-PSVINNETRO TK <mark>P</mark> DISIRSIKAH	<pre>(YRYSFQTRE - I IRSHCG (YRYSFTTEN - IVRSNCG QYEYSFKTRH - ITRKDCG (HVFSAVLSN - LTSSMCG</pre>	SKKIIYNARN - EFGNI - EEKSTIYIS SKKIMYHARN - EFGNV - EGYSMIYIS SRNLTLVTSN - TIGSM - EINAKIDVK SKLTFKASN - KFGKV - STSAVINVS
	Alr30-F-1_D2/1-97 Alr31-F-1_D2/1-99 Alr33-F-1_D2/1-93	<mark>GGP</mark> SIISNLSSTY <mark>T</mark> VE GGPSIC <mark>G</mark> AL <mark>P</mark> SPVTYL GGPFVCNKERK <u>-</u> IV	ENTTN <mark>YNVP</mark> IILQ <mark>G</mark> KSRDQ <mark>T</mark> TLSVFLCG I <u>R</u> TN <mark>PL</mark> RVKDIICS	HPK - PQVTWEFAGN HPT - PEVIWKVDNF NPK - PMVTWYVDNN	KN <mark>V</mark> R <mark>M</mark> EA I AEKKRI F - L N <mark>G</mark> T VKE I DEETEL L - IS <mark>GG</mark> DVLKKQLT	<pre>KYKYILTIPI-ITREMNG .FKYTLDFKNYLQPKKCS KYSVRKIFTT-QDVSLCG</pre>	)KELSYKIVG - ALSEN - TLVGSIKLT KVLSYKAVGYNDNEV - VGSTKLNVT SVKVSLMARGHSEEEQ - VFQSEIQYG KRISYVATG - RNGI <mark>S</mark> - LNR <mark>S</mark> TLLVF
	Alr35-F-1_D2/1-96 Alr36-F-1_D2/1-100 Alr37-F-2_D2/3-94	<mark>GGPEFCNLKPP</mark> ETLTV GGPDDCGD <mark>S</mark> LKSMYNV EKPYRCGHRTLQQ	ISHER <mark>P</mark> RVRVSVCG E <mark>E</mark> FEKLLVTKEICG ISSSSFQVK <u>E</u> TICG	NPE - PQITWEYLGN HPK - PTVRWKFQRN QRI - PSLVWYLD <mark>G</mark> K	N - L K S <mark>T</mark> I I K <mark>G</mark> KK <mark>P</mark> KE RF YS - <mark>P</mark> SAS <mark>S</mark> QL I NNSTRI L - IG - NGTA <u>-</u> RE I EF	EFI <mark>FETLLPQ-ITPP</mark> NCG .YR <mark>YS</mark> FEWRN-ITRANCG KHK <mark>Y</mark> VFTKTF-KDMTACG	SETVRYVAAG - YGREL - TSSTVIKVS SSLTYKATG - HGRGV - SDRTTLNVK SRKLLFSASG - YGESL - SENAWLNVV SRNLSYVITS - FDGYTATYYATLSEN
Alr D3	Air38-F-1_D2/1-95 Air2-F-1_D3/1-94 Air30-F-1_D3/1-94 V-frame residues	- SPTFCDTAVK - VIEF	EEVASAMFTTHVCG AEVASAVFTTHVCG	NPM - PTVSMEFGGC	S-VAVTHNTAEKD	(YK <mark>YTANLT<u>S</u>YLK<mark>P</mark>SRC<mark>G</mark></mark>	SLLLRYVASG - HDGEI - TGSTKLIEK RELIATAEN - SIGEK - TSKVVLKYK KKLIITAKN - NKGEG - TRRIVLKYK Cys <sup>104</sup>
	(sensu Cannon <i>et al.</i> 2002 I-set motifs (Wang 2013)	2) Pro	Gly <sup>16</sup> Cys <sup>23</sup>	trp³⁵	(or Lys/His)	Leu <sup>89</sup> Asp <sup>36</sup>	Tyr <sup>102</sup> Cys Tyr Asn Pho-X-Pho-X
		ix. Alr-specific	iii. ii. Alr-spe	v. iv.		viii. Alr-spe	i. i. vi. vii. cific
		conserved cysteine	consen cystei			conser cystei	

# Fig. S16. Multiple sequence alignment of I-set Ig domains and AIr domains 2 and 3.

Alr domain 2 and 3 sequences were aligned to I-set Ig domains from pfam. Residues in the alignment are highlighted by sequence conservation and chemical property with CLUSTALX colors as implemented in Jalview. The positions of conserved V-frame residues are shown above and below the alignment with gray background. Motifs common to I-set domains are also indicated. The position of invariant cysteine residues is shown in red lettering. Note that domain 3 is the most membrane-proximal Ig domain in Alr2 and Alr30, hence the conserved cysteine appears there and not in domain 2.

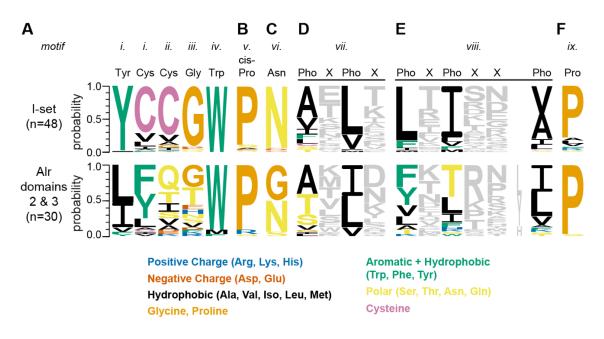
Distance between O(i) and N(i+3)

Distance between  $C^{a}(i)$  and  $C^{a}(i+3)$ 





Left graph shows distance between the alpha carbons of the first and fourth residues in the turn. In a beta-turn, the distance between these two atoms is less than 7.0 Å. Turns <7.0 Å are shown in black bars. Those >7.0Å are showing with gray. Right graph shows distance between the oxygen atom of the first residue and the nitrogen atom of the fourth residue. A distance of <3.5Å is considered small enough for hydrogen bonding to occur. Black bars indicate distances <3.5Å. Gray bars indicate distances >3.5Å. Gene names are color-coded according ot whether they are *bona fide* (blue) or putative (orange) genes.



# Fig. S18. Sequence logo of I-set-specific motifs in classical I-set domains and Alr domains 2 and 3.

Sequences of Alr domains 2 and 3 and I-set domains from the pfam I-set sequence profile (pf07679) were aligned in MAFFT. Sequence logos for then created to represent the motifs identified by Wang (2013) for the Alr or pfam sequences.

(A) Motifs *i*, *ii*, *iii*, and *iv*, which are discussed in the main text because they are also part of the V-frame as described by Cannon et al (2002).

(B) Motif v is a conserved proline in the BC loop, and motif vi is a conserved asparagine in the FG loop. These two residues form a hydrogen bond that stabilizes the BC and FG loops in a closed position. In domains 2 and 3, we found the conserved proline residue in 28/31 sequences but the asparagine was only present in 13/31 sequences (see also Figure S9). Thus, motif v is present in domains 2 and 3, but motif vi and the structural motif it forms with motif v do not appear to be a common feature of these domains.

(C) Motif *vii* is a Pho-X-Pho-X pattern of amino acids (where Pho represents a hydrophobic residue), located approximately 10-12 residues downstream of motif *vi*. It is found in beta-strand G and denotes the C-terminal end of an I-set (and also V-set) domain. This pattern was found in 30/31 of domain 2 and 3 sequences (See also Figure S9).

(D) Motif *viii* is a set of hydrophobic and hydrophilic residues, Pho-X-Pho-X-X-Pho, located at the bottom of beta-strand E. The last two hydrophobic residues typically contact the tyrosine in the tyrosine corner, while the two consecutive hydrophilic residues between them form a beta bulge. In our alignments, 26/31 domain 2 and 3 sequences had this motif, although the first and second hydrophobic residues often had polar or aromatic side chains (see also Figure S9). As noted above, the tyrosine corner is not present in domains 2 and 3. However, a beta bulge was predicted to occur at the end of the E strand in 28/31 structural models. Motif *viii* is therefore present in most

domains, but the structural consequences of this motif are likely to differ from traditional I-set domains.

(E) Motif *viiii* is a proline ~23-26 residues upstream of the B-strand cysteine. This proline defines the beginning of an I-set domain and, in domains 2-3, it was found in 30/31 sequences (see also Figure S9).

# Removed

LAR_DROME/418-503 TENA_CHICK/1495-1571	10 - SAPRNVQ	20 VRTLSS-ST-MVITWEP	90 40	50 60 70	80 ASWNS	90 10 QMV	0     110     10     10     1	1 VSEL-	20 130 TPHAIYTVRVQAYT	140 SM <mark>G</mark>	150 160 170 1	30 <u>190 200 210 220 23</u> 1 1 1 1 1 1 1
KALM_CHICK/1495-15/1 KALM_CHICK/544-641 LAR_DROME/710-800	- AKDENI SASE.	- TVOEGN - TT - GHESWKT			ESPONSI PNSTTSOS		- AD HVVI T		PRSMINEL EVOVI T	T <mark>66</mark>	F <mark>GP</mark> A-T	
LAR_DRUME/710-800 PTP99_DROME/172-259 IL7RA_HUMAN/130-218	-SKPONLT	ILDVSA-NS-ITMSWHP		AIAGYHVEHIHDNO-TGV	EIVKNSRN	s <mark>v</mark>	-ETLIHEE	LONL-	RPYTDYRVIVKAFT	TKN	EGEP-S	
TIE2_HUMAN/445-529 TIE1_HUMAN/447-533	<mark>P</mark> K <mark>PL</mark> NA <mark>P</mark> N·	VID <mark>TG</mark> H-NF-AVINISS	5E <mark>P</mark> YF <mark>G</mark> DG-	<mark>PI</mark> KSKKLL <mark>Y</mark> K <mark>P</mark> VNH-YE	AWQHI	Q <mark>V</mark> T	NEIVT	LNYL-	EPRTEYELCVQLVR-P	R <mark>G</mark> -E <mark>GG</mark>	E <mark>G</mark>	-нббб
PTPRK_MOUSE/291-376	PRPIAPPQ-	LLGVGP-TY-LLIQLNA	NSIIGDG-	<mark>PIILKEVEY</mark> RMTSG	SWTET	H <mark>A</mark>	-VNAPTYK	LWHL-	DPDTEYEIRVLLTR-P	PG-EGG	T <mark>G</mark> L <mark>P</mark>	G
PTP10_DROME/218-301 PTP10_DROME/124-205	-DPPSNLS	VQVRSG-KN-AIILWSP	QG-	S <mark>Y</mark> TAFKIKVLGLSE-AS	SSYNRTF	Q <mark>V</mark> N	-DN TFQHS	VKEL-	TPGATYQVQAYTIY		D <mark>G</mark> KE-S	
FINC_BOVIN/1635-1713 FINC_BOVIN/1176-1257	-SPPTNLH	LEANPD-TGVLTVSWER	RTP-	DITGYRITTTPTNG-QQG	YSLEE	v <mark>v</mark> H	-ADQSSCT	FENL-	SPGLEYNVSVYTVK	DDK	ES	
FINC_BOVIN/1907-1984 FINC_RAT/1267-1346	-PQLTDLS	FVDITD-SS-IGLRWT	SS-	TIIGYRITVVAAGE-GI	<mark>P</mark> IFED	F <mark>V</mark> D	-SSVGYYT	VT <mark>GL</mark> -	E <mark>PGIDY</mark> DISVITLI	N <mark>GG</mark>	ES	
FINC_BOVIN/909-987 FINC_BOVIN/1360-1439	-PPPTDLR	FTNVGP-DT-MRVTWAP	ss-	IELTNLLVRYSPVKN-EE	DVAEL	s <mark>1</mark> s	-PSDNAVV	LTNL-	L <mark>PG</mark> TE <mark>Y</mark> LVSVS <mark>S</mark> VY	EQH	ES	
FINC_BOVIN/1815-1893 TENA_CHICK/683-767	-PAPEGLK	FKSVRE-TS-VOVEWDP	SI-	SFDGWELVFRNMOK-KD	DNGDITS	s <mark>L</mark> K	-RPETSYM	OPGL-	A <mark>PG</mark> QO <mark>YNVSLHIVK</mark>	NNT	R <mark>GPG</mark> -LS	
TENA_CHICK/775-853 FINC_BOVIN/1451-1529	-DS <mark>P</mark> S <mark>GI</mark> D	FSDITA-NS-FTVHWIA	4RA-	TIT <mark>GY</mark> RIR <mark>H</mark> HPENM- <mark>GG</mark>	R <mark>P</mark> RED	R <mark>V</mark> <mark>P</mark>	-PSRNSIT	LTNL-	NPGTEYVVSIVALN	SKE	ES	
TENA_CHICK/1318-1395 TENA_CHICK/594-671	-SPPTELT	VTNVTD-KT-VNLEWKH	EN-	LVNEYLVTYVPTSS-GG	LDLQF	T <mark>V</mark> <mark>P</mark>	-GNQTSAT	IHEL-	EPGVEYFIRVFAIL	KNK	KS	
TENA_CHICK/957-1033 TENA_CHICK/865-945	-DA <mark>P</mark> RNLK	-RVSQTD-NS-ITLEWKN	IHA-	NIDNYRIKFAPISGG	DHTEL	T <mark>V</mark> <mark>P</mark>	-KGNQATTRAT	LT <mark>GL</mark> -	R <mark>PGTEYGIGVTAV</mark> R	QDR	ES	
FINC_BOVIN/1541-1622 FINC_BOVIN/1996-2074	-DAPSNLR	FLA <mark>T</mark> T <mark>P</mark> -NS-LLV <mark>S</mark> WQ <mark>P</mark>	RA-	R <mark>ITGY</mark> IIK <mark>Y</mark> EK <mark>PG</mark> S	<mark>PP</mark> REV	V <mark>P</mark> R <mark>P</mark>	R <mark>PG</mark> VTE <mark>A</mark> T	IT <mark>GL</mark> -	E <mark>PG</mark> TE <mark>Y</mark> TIQVIALK	NNQ	кѕ	
TENA_CHICK/1407-1483 PTP69_DROME/333-425	DPIFIPKVE	TTGSTA-ST-ITIGWNP	PDLID-	YIQYYELIVSESGE	-VPKVIEE	AIY00N	-SRNLPYM	FDKL-	KTATDYEFRVRACS-D	DL-TKT	C <mark>GP</mark> W-S	
FINC_BOVIN/813-890 FINC_BOVIN/722-800	-VVATSES	VTEITA-SS-FVV <mark>S</mark> WVS	5ASD-	TVSGFRVEYELSEE-GD	E <mark>P</mark> QYL	D <mark>L</mark> <mark>P</mark>	-STATSVN	IPDL-	LPGRKYTVNVYEIS	EE <mark>G</mark>	EQ	
PTPRB_HUMAN/1356-143 PTP10_DROME/406-485	-LPVRNLR	SINDDKT-NT-MIITWEA	DPAS-	TODEYRIVYHELETFNG	DTS	TLT	-TDRTRFT	LESL-	LPGRNYSLSVOAVS	KKM	ES	
PTPRB_HUMAN/468-543 PTP10_DROME/313-394	-LR <mark>P</mark> LNVTFD	RDFITS-NS-FRVLWEA	A <mark>P</mark> K <mark>G</mark> IS-	EFDK <mark>Y</mark> QVSVATTRR	Qs	T <mark>V</mark> <mark>P</mark>	-RSNEPVAFF	FDFRDIA	E <mark>PG</mark> KTFNVIVK <mark>T</mark> VS	<mark>G</mark> KV	TS	
PTPRB_HUMAN/113-192 PTPRB_HUMAN/644-725	-ssvsgvtv	NNSGRN-DY-LSVSWLL	PG-	DVDNYEVTLSHDGK	VVQSL	VIA	-KSVRECS	FSSL-	TPGRLYTVTITTRS	<mark>6</mark> КҮ	ENHSFS	
PTPRB_HUMAN/732-808 PTPRB_HUMAN/908-984	-SAVKNIHI	<mark>SP</mark> NGAT-DS-LTVNWTP	GGGG	DVDSYTVSAFRHSQ	KVDSQ	T <mark>I</mark> <mark>P</mark>	-KHVFEHT	FHRL-	EAGEQYQIMIASVS	<mark>G</mark> SL	КК	
PTPRB_HUMAN/996-1074 PTPRB_HUMAN/555-632	-AQVTDLHV	ANQGMT-SS-LFTNWTQ	<u>0</u> Q <mark>G</mark> -	DVEFYQVLLIHENV	VIKNE	s <mark>1</mark> s	-SETSRYS	FHSL-	K <mark>SG</mark> SLYSVVVTTVS	<mark>GG</mark> I	SS	
PTPRB_HUMAN/1086-116. PTPRB_HUMAN/1174-125 PTPRB_HUMAN/1262-134	0 - ASVSHLRG	<mark>S</mark> NRNTT-DS-LWFNWS <mark>P</mark>	S <mark>G</mark> -	DFDFYELILYN <mark>P</mark> NG	TKKEN	WKD	-KDLTEWR	FQGL-	V <mark>PG</mark> RKYVLWVVTHS	<mark>G</mark> DL	S	
PTP10_DROME/865-939	-EPITQLH	ATNITD-TE-ISLRWDL	KG-	EYNDFDIAYLTADN	LLAQN	M <mark>T</mark> T	-RNETT	ISDL-	RPHRNYTETVVVRS	<mark>G</mark> TE	SS	
PTP10_DROME/584-661 NGCA_CHICK/702-794	- RNPGGVH	GEGNET-GN-LVTTWEP		PWARYRVOWRPLEE-PGGGGPS	GGEPWAES	TVD	APPVV	VGGL -	PPESPEOTRVOAVN	<mark>GAG</mark>	KGPF-A	
CNTN2_CHICK/604-692 MYPC2_CHICK/632-717	-DPPOSVR	VTSVGE-DW-AVLSWEA	PPFDGGM-	PITGYLMERKKKGS-MRW	MKLNF	EVFP	DTTYE	STKM-	IEGVFYEMRVFAVN	AIG	VS0 <mark>P</mark> -S	
MPSF_CHICK/700-785 MPSF_CHICK/802-887	-GPAYDLT	VCEVRN-TS-LVLLWKA	APVYE <mark>G</mark> KS-	PITGYLVDYKEVDT-EDW	ITAN	EKP	-TSHRYEK	VTDL-	HOGHTYVEKVRAVN	DA <mark>G</mark>	V <mark>G</mark> KS-S	
MPSF_CHICK/372-457 MPSF_CHICK/500-585	- <mark>GPP</mark> TN <mark>V</mark> H	ASEISK-TY-VVLSWDP	PVPRGR-	EPLTYFIEKSMVGS-GSW	QRVNA	0 <mark>V</mark> A	-VKSPRYA	VFDL-	AEGKPYVFRVLSAN	KH <mark>G</mark>	ISD <mark>P</mark> -S	
MYPC2_CHICK/925-1006 F4H3U7_CELFA/652-733	-TTPGTPV	- ATGVTT-VG-ASLSWAA	S TDAGS-	GVAGYELYRVOGTT	0	TLVGT	-TTAAAYI	LRDL-	TPGTAYSYVVKAKD	VA <mark>G</mark>	NVSAA-S	
MYLK_CHICK/1319-1403 MPSF_CHICK/601-684 NRCAM CHICK/624-709	-SAPGRVV	ATRNTK-TS-VVVOWDK		NLYGYYIDYSVVGS-NOW	EPAN	HK <mark>P</mark>	-INYNREV	VHGL-	ETGEOYIFRVKAVN	AVG	FSEN-S	
NRCAM_CHICK/624-709 LAR_DROME/323-404 NCAM1_BOVIN/611-691	- TAPTDVQ	ISEVTA-TS-VRLEWSY	/-K <mark>G</mark> PE	DLQYYVIQYKPKNA-NQ	S	E <mark>I</mark> S	GIITMYYV	VRAL-	SPYTEYEFYVIAVN	NI <mark>G</mark>	RGPP-S	
NCAM1_BOVIN/611-691 NCAM1_BOVIN/510-597 FINC BOVIN/610-692	-SS <mark>P</mark> S	IDQVEPYSSTAQVQFDE	PEATGG-	VPILKYKAEWRAMGE-EVWHS	KWYDAKEA	SM	EGIVT	IVGL-	KPETTYAVRLAALN	<mark>G</mark> KG	L <mark>G</mark> EI-S	
PTPRB_HUMAN/24-103	PERCNFTL	AESKASS-HS-VSIQWRI	L <mark>G</mark> -	SPCNFSLIYSSDTL-GAAL	C <mark>P</mark> TF	R1D	-NTTYGCN	LQDL-	QAGTIYNFRIISLD	EER	T	
TIE2_HUMAN/640-724 PTP69_DROME/237-321 TENA CHICK/1228-1306	POVSID	FAKAVGA-NK-IYLNWTV	/NDGND-	PIOKFFITLOEAGT-PTF	TYHKD	F <mark>I</mark> N	-GSHTSYI	LDHF-	KPNTTYFLRIVGKN	SI <mark>G</mark>	NGOP-T	
TENA_CHICK/1228-1300 TENA_CHICK/1046-1124 TENA_HUMAN/1257-1335	-PELGNLS	VSETGW-DG-FOLTWTA	4D <mark>G</mark> -	AVENEVIOVOOSDN-PE	ETWNI	TVP	-GGOHSVN	VTGL-	KANTPYNVTLYGVI	R <mark>G</mark> Y	RT	
TENA_CHICK/1137-1215 TENA HUMAN/1530-1608	-PEVGELT	VSDITP-ES-ENLSWTT	N <mark>G</mark> -	DFDAFTIEIIDSNR-LL	EPMEF	NIS	-GNSRTAH	ISGL-	SPSTDFIVYLYGIS	H <mark>G</mark> F	RT	
EPHB2_CHICK/327-422 EPHA3 CHICK/326-421	-SAPOAV	ISSVNE-TS-LMLEWTP	RDSGGF	REDLVYNIICKSCGS-GRG	-ACTRCGD	NVOFA-PROL	GLTEPRIY	ISDL-	LAHTOYTFEIOAVN	GVT	DOS <mark>P</mark> F-S	
EPHA1_HUMAN/334-431 EPHA2_HUMAN/330-419	-SAPRNLS		PADTGG	R QDVR <mark>Y</mark> SVRCSQCQ <mark>G</mark> TAQD <mark>GG</mark>	- PCQPCGV	GVHFS-PGAR	GLTT <mark>P</mark> AVH	VN <mark>G</mark> L-	EPYANYTENVEAQN	<mark>G</mark> VS	<mark>G</mark> L <mark>G</mark> SS	
7LESS_DROVI/1918-199 TIE2 HUMAN/544-626	7 YAPLPPLQ	LIELNA-YG-MTLAWPG	- T P D -	ALSSLTLECOSLR	EQLQF	NVA	-GNHTQMR	LAPL-	QPKTRYSCRLALAY	AAT	<mark>PGAP</mark> -I	
KALM_CHICK/178-269 LAR DROME/516-598		FIELQS-GD-LEVKWSS	KFNISIE <mark>P</mark>		H <mark>P</mark> SEDDATNWQ	T <mark>V</mark> AQ	- TT DERVQ	LSDI-		VH <mark>G</mark>		
PTPRZ_HUMAN/313-401 EPHA2_HUMAN/437-519	-SEPENVO	-ADPENY-TS-LLVTWER	RPRVVYDT-	MIEKFAVLYOOLDG-ED	OTKHE	FLTD	GYODLGAI	LNNL-	LPNMSYVLOIVAIC	TNGL	Y <mark>G</mark> KY-S	
TIE1_HUMAN/645-729 ITB4 HUMAN/1529-1612	-PAPRHLH	AQALSD-SE-IQLTWKH	IPEALPG-	<mark>PI</mark> SK <mark>Y</mark> VVEVQVA <mark>GG</mark> -A <mark>G</mark>	D <mark>P</mark> LWI	D <mark>V</mark> D	RPEETSTI	IR <mark>GL</mark> -	NASTRYLFRMRASI-Q	Q <mark>G</mark>	L <mark>G</mark> DW-S	
ITB4_HUMAN/1221-1310 ITB4_HUMAN/1642-1728	-SE <mark>PGRL</mark> A	FNVVSS-TV-TQLSWAE	EPAETNG-	EITAYEVCYGLVND-DNRPIG-	<mark>Р</mark> МККV <b>Р</b> АТАF	LVD RVDGD	NPKNRMLL	IENL-	RESQ <mark>PYRYTVKARN</mark>	<mark>G</mark> AG		
PTP10_DROME/959-1044 EPHB2 CHICK/438-521	-GRVERFH	<mark>P</mark> TD <mark>VQP</mark> - SE - INFEWSL OVSRTV - DS - ITL <b>S</b> WSC	PSSEANG- )PDOPNG-	VIRQESIAYTNINNLTDAGM VILDYELOYYEKNL-SELNS	Q T	D <mark>F</mark> E AVK	-SEEAFGV	IKNL- VONL-	K <mark>PG</mark> ET <mark>YVFKIQA</mark> KT KAGTIYVFOVRART	AI <mark>G</mark>		
EPHA4_MOUSE/442-525 LAR_DROME/911-995	-SSIALVQ	AKEVTR-YS-VALAWLE	PPDRPNG-	VILEYEVKYYEKDQ-NERSY	R	I <mark>V</mark> R	-TAARNTD	IKGL-	NPLTSYVFHVRART	AA <mark>G</mark>	Y <mark>G</mark> DF-S	
7LESS DROME/1800-189	1 - SPPRNFS	VRVLSP-RE-LEVSWLP	PEQLRS-	ESVYYTLHWOOELD-GENVODR	REWEAHER	RL	-ETAGTHR	LTGI-	KPGSGYSLWVQAHA	T <mark>P</mark> T	KSNS-S	
E1JJF8_DROME/918-100 UF0_HUMAN/335-418	-GPPENIS	ATRNG-SQ-AFVHWQE	PRAPLQG-	TLLGYRLAYOGODT-PE	VLMDI	GL	-RQEVTLE	LQGD-	GSVSNLTVCVAAYT	AA <mark>G</mark>	DGPW-S	
ITB4_HUMAN/1128-1208 E1JJF8_DROME/817-905	- GAPQNPN	AKAAGS-RK-IHFNWLP MROITSSTS-GYMAWTP	SG- VS-EESVRG-	K <mark>P</mark> MGYRVKYWIQGD-SESEA HFKGYKIOTWTENE-GEE	н <mark>G</mark> LREI	L <mark>L</mark> D HVK	- SK V <mark>P</mark> SVE - GD THNAL	LTNL-	Y <mark>P</mark> YCD <mark>Y</mark> EMKVCAYG KPDSKNYARILAYN	AQ <mark>G</mark>	EGPY-S NGPP-S	
CNTN2_CHICK/809-896 CNTN1_CHICK/801-884	-EVPTDVS	VKVLSS-SE-ISVSWH	EK-	SVEGYOIRYWAAHD-KEA	AAORV	o <mark>v</mark> s	-NOEYSTK	LENL-	KPNTRYHIDVSAFN	SA <mark>G</mark>	Y <mark>GPP</mark> -S	
TIE1_HUMAN/547-632 L1CAM_HUMAN/813-907	QAIPELEG	IEILNS-SA-VLVKWRP	VD-LAQVKG-	HLRGYNVTYWREGS-QRK	HSKRHIHKDHV	V <mark>V</mark> <mark>P</mark>	-ANTTSVI	LSGL-	RPYSSYHLEVQAFN	<mark>GRG</mark>	S <mark>GP</mark> A-S	
ALr01.1_ECS/235-359 ALr02.1_ECS/324-429	-FTPGKVQNL	KSSRKD-KC-II <mark>TTW</mark> KN	VDT <mark>G</mark> N-	CEVWYTVKYGGEE	LLHEQ	N <mark>T</mark> S	-SGKVGASFO	CN <mark>S</mark> D	KVPKVNITRIVAVS	NDT	FKQE <mark>G</mark> EE-EIVTV	VLTT <mark>P</mark> STTTTTTTTTTSTRKSKTTTNKS <mark>GG</mark> QVITNN <mark>GNG</mark> TNI
ALr03.1_ECS/243-364 ALr04.1_ECS/240-357	- FVPAKVMGGN-	FYERDNC-TYVTWKR	RERT <mark>G</mark> N-	CRVMYYLQFGNGA		к <mark>v</mark> N	-TLGTMYKKO	CNDA	VLMQVDSVTIWGKY	<mark>G</mark> LK	QGEK-FTLVK	SLTA <mark>P</mark> TTVSTTTKSESS <mark>G</mark> AAGDDDDDDDKT
ALr06.1_ECS/238-337 ALr07.1_ECS/360-465	-FKPSPVSINS	MYRINE-SC-VYMGWLG	ESAEN-	CSVKYYFOFDGEH		SRHE	-ISAMNEVHO	c <mark>g</mark>	- LONARSVVFWASY	KNI	I <mark>G</mark> KK-TNALL	SSTR <mark>PP</mark> N <mark>PP</mark> KKDDEDNK <mark>G</mark> GQCANKGQCANK
ALr08.1_ECSb/583-688 ALr08.1_ECSa/251-363	-FTPSTVSIKS	LYSTSL-NC-IHATWTR	₹EDT <mark>G</mark> N-	CVVNYHLQFNSRN		D <mark>I</mark> YS	-TSKSYFTIC	CN	LRSRPAFDTIWASY	K <mark>G</mark> L	L <mark>G</mark> KK-HSSST	S <mark>PATPAP</mark> ETKEISADESKLTIKLI T <mark>G</mark> TV <mark>P</mark> KTTTTTTKVVIRT <mark>P</mark> SCLSYYLTNCT
ALr09.1_ECS/240-350 ALr11.1_ECS/235-345	- F S <mark>P</mark> Q R I R K V K ·	FYKDNNC-IN <mark>GTW</mark> TS	EAT <mark>G</mark> N-	CALNYHLQFAERS		D <mark>I</mark> LN	-TSDTHYAVO	CN	-IFNVSSVVIWASY	KNN	YGQK-AKVNI	SLTTPAPSTKVTDTCKRPILLKVKSLDTKK
ALr12A.1_ECS/240-349 ALr12B.1_ECS/242-351	-ESPOKTONAV	FYKDNDC-TNGTWTC	FAT <mark>G</mark> N-			Y <mark>T</mark> FY	-STHTY	CN	- VI NASYVETWASY	KNK	Y <mark>G</mark> FK-TKTNA	SLTT <mark>PAP</mark> -TNATDTCTC <mark>P</mark> KLLEVKRLDAKK
ALr15.1_ECS/247-346 ALr16.1_ECS/243-366	- FKPSKITLTT	SYRYNA-SC-VYLTWHK	(EDT <mark>G</mark> N-	CLLTYHLKFDDKD		D <mark>V</mark> YS	-TFNTNFNLQ	сн	-SSCAASASVWASY	K <mark>G</mark> N	AGYK-NSINL	NAERVAQRKDNNLK
ALr18.1_ECS/243-359 ALr19.1_ECS/242-354	- FVPSMVSMVS	LYRHNA-SC-VRVTWDA	EDT <mark>G</mark> R-	CNVSYHLOFTGRE		TIYN	-SSNRYFTLO	CN	-SSDVDTVIIWASY	K <mark>G</mark> K	NGWK-LASRI	STTTSTPDITSTLSILTTNSKKLVVTSDNNQCTSN STTTSSPYFTSTTNIATTNAGKSVPRTCDCQ
ALr21.1_ECS/244-353 ALr23.1_ECS/242-371	-FVPSPVSMIS	LSSYNG-TC-VRVTWDA	AEDT <mark>G</mark> K-	CILNYHVWFSGRE		I <mark>I</mark> YN	-TSNTYFTLC	CN	-ATDVKNVTIWASYNF	RDVK <mark>G</mark> N	FTTT-SSPDITST	HLITLR <mark>P</mark> DITILSSVTTTSKDDWHCASI TNSTTTLSPDITST <mark>P</mark> NIATTNAVKCVQRTCKLKCDCQATSN
ALr27.1_ECS/248-344 ALr28.1_ECS/229-331	-FSPSKVTDF-	SFAIKD-NC-QIFQWST	LNSGR-	CGVFYEIQILDNKK	RILSQ	STTQ	-PFANFLSFC	CYAE-	YTHNNVSAGIRAIY	DNK	Y <mark>G</mark> NW-SSVK <mark>P</mark>	TVGIPQSKDNASF
ALr30.1_ECS/321-423 ALr30.3_ECS/321-445 ALr31.1_ECS/236-333	-LS <mark>P</mark> VQIN	-STMRKVG-SC-IDTNWSA	APNTGE-	CDVSYKVDYLDSNN	VIVHS	KVFH	-NKELKTESC	CD <mark>T</mark> K	VVSNTLSVRVTVVS	NDTRVI	EPTNESILCGRITLVKYFCEFWSTRGNKFSPPKVL	DKPTDNDSADLKDKVSSV
	ALISIK	BLKKKM-DC-LFTEWTP	KHVDK-		D	vewken	-II (NEALFO	CSNV-			YGDW-YIHAL	STNDQKKDHN SYET <mark>P</mark> LNI <mark>G</mark> QQSNILR <mark>PTP</mark> VS
Alr33.1_ECS/230-342	-FK <mark>P</mark> DQ <mark>V</mark> QIL											
	-FK <mark>P</mark> DQ <mark>V</mark> QIL- -SLPK <mark>PVENFK·</mark> -FT <mark>P</mark> ENVKVT	<mark>YYPMGG</mark> -NC-FKF <mark>T</mark> WLA EAYLKE-QC-VTVRFTT	AQNT <mark>G</mark> L- TLDV <mark>G</mark> T-	CKTH <mark>FELQL</mark> LLKNN	KVKRQ QLV <mark>G</mark> S	S <mark>I</mark> S <mark>P</mark> S <mark>A</mark> AD	-MTTGTFIHO	CE <mark>V</mark> DS <mark>A</mark> - C <mark>GITAS</mark> -	SLKKI <mark>Y</mark> KAKIR <mark>S</mark> IY TVKARARS	QD <mark>G</mark>	ATRLE <mark>G</mark> RW-SLLEL VGQW-SAYHI	EKSGODNKDTNNNID

Fig. S19. Amino acid sequence removed from the ECS prior to structure prediction.

Alr1 ECS-trimmed Alr2 ECS-trimmed Alr3 ECS-trimmed Alr4 ECS-trimmed |--A---| |--B--| |--C-| |C'| |-E-| |-F---| |G-| FVPAKV<mark>MGGNFYER</mark>DN<mark>CTYVTWK</mark>RERTGNCRV<mark>MYYLQF</mark>GNGA<mark>KVNT</mark>LGT<mark>MYKKC</mark>NDAVLMQVD<mark>SVTIWGKY</mark>GLKQGEK<mark>FTLV</mark>K EEEEEEEETTEEEEEEE TTTT EEEEETTT EEEE EEEEETTGGGGGG EEEEEEETTEE EEEE Alr6 ECS-trimmed Alr7 ECS-trimmed |--A---| |--B--| |--C-| C' |-E-| |--F---| |G-| FKPSPV<mark>SINSMYRI</mark>NES<mark>CVYMGWL</mark>GESAENCSV<mark>KYYFQF</mark>DGEHSR<mark>HEI</mark>SAM<mark>NFVHC</mark>GLQNAR<mark>SVVFWASY</mark>KNIIGKK<mark>INAL</mark>L Alr8 ECSa-trimmed Alr9 ECS-trimmed |--A---| |--B--| |-C--| |C'| |-E-| |--F---| |--G---| FSPRKI<mark>RKVIFYKE</mark>ND<mark>CINGTWT</mark>SEGTGNCPL<mark>TYHIHE</mark>NKENN<mark>IFNT</mark>SDT<mark>HYAVC</mark>NISNVS<mark>SVVIWASY</mark>RN<mark>KY</mark>GQR<mark>TKVN</mark>I Alr11 ECS-trimmed Alr12A ECS-trimmed AIFIZA ECS-EFINNNEG |--A---| |--B--| |--C-| |C'| |-E-| |--F---| |---G---| FSPQKI<mark>QNAIFYKD</mark>ND<mark>CINGIWT</mark>REATGNCVL<mark>NYHLQF</mark>EGGRY<mark>IFNT</mark>THT<mark>YYAVC</mark>NVLNAS<mark>YVFNWASY</mark>KNKYGEK<mark>IKIN</mark>A Alr12B ECS-trimmed |--A---| |--B--| |--C-| |C'| |-E-| |--F---| |---G---| FSPQKI<mark>QNAVFYKD</mark>ND<mark>CINGIWT</mark>CEATGNCVL<mark>NYHLQF</mark>EGGRY<mark>IFYS</mark>THT<mark>YYAVC</mark>NVLNAS<mark>YVFIWASY</mark>KN<mark>KY</mark>GEK<mark>IKIN</mark>A Alr15 ECS-trimmed |---A---| |--B---| |---C-----| |--C'---| |E-| |---F----| |G-| YT<mark>PQNVSYKAD</mark>SNG<mark>CYTLTWSI</mark>HHLGIKC<mark>VKMYELKALFDD</mark>ES<mark>LLKNTTILT</mark>TLN<mark>NNFM</mark>CYPS<mark>ILYVKVRTVS</mark>IWNAKSNW<mark>VMQE</mark>I Alr16 ECS-trimmed |--A---| |--B--| |--C-| |C'| |E-| |--F---| |---G---| FKPSKI<mark>TLTTSYRY</mark>NAS<mark>CVYLTWH</mark>KEDTGNCLL<mark>TYHLKF</mark>DDKDD<mark>VYST</mark>FNT<mark>NFNLC</mark>HSSCAA <mark>SASVWASY</mark>KG<mark>NAGYKNSIN</mark>L Alr18 ECS-trimmed Alr21 ECS-trimmed |--A---| |--B--| |--C-| |C'| |E-| |--F---| |---G---| FTPSKV<mark>RITSSYRH</mark>NAS<mark>CVYLNWY</mark>REDTGNCLL<mark>EYHIQF</mark>NNIND<mark>VYNT</mark>SKT<mark>NVDI</mark>CHSPSAS<mark>SASIWASY</mark>KG<mark>ISGGKVDII</mark>L Alr23 ECS-trimmed |--B--| |--C---| |C'| |-E-| |--F---| |---G---| |---A---|

	A	B	<u>I</u> <u>I</u>	c	C'	-E-	F	-   G		
FVPSPV	SMISLSS EEEEEEEE	(NGT <mark>CVRVTW</mark> ETTTEEEEEE	<mark>D</mark> AEDTGKCI <mark>L</mark> E TTTT E	<mark>nyhvwfs</mark> gi eeeeeett	REI <mark>IYNT</mark> SN F EEEE	P <mark>YFTLC</mark> NAT EEEEETTT	IDVK <mark>NVTIWA</mark> ITT EEEEEI	<mark>5¥</mark> NRD <mark>VKGNFTTTS</mark> EETTTEEEEEEEE	SPDI	
 YT <mark>PIDI</mark>	KFEKKS II	B  CYNLSWSVH	PLAKPC <mark>VKEY</mark>	QLVVNQ ID	KAPLNTTTK!	rs <mark>nylic</mark> se	ks <mark>iqsckvrt</mark> i	-   G-  L <mark>C</mark> KSNQESDW <mark>VTAR</mark> ZETTT B EEEE	I	
FSPSKV	TDFSFAIF	B  Adn <mark>cqifqws</mark>	TLNSGRCGV <mark>F</mark>	YEIQIL DNH	KKR <mark>ILSQST</mark>	IQPFAN <mark>FLS</mark>	<mark>sf</mark> cyaeythni	F  NV <mark>SAGIRAIYDNKY</mark> TTEEEEEEETTEE	GNW <mark>SSV</mark> KP	
LTPESV	KNVNSTEF	B  NDG <mark>CMYTTWE</mark>	kvntgeca <mark>vs</mark>	YRIDYF <mark>DSI</mark>	RGL <mark>VLFSQNI</mark>	KSEGIY <mark>TAS</mark>	SMCDEIIISKV	F  /N <mark>NLRIIAIP</mark> NSIE EEEEEEE TTT	rkvhg <mark>ngtivk</mark> i	
LSPVQI	-A  N <mark>STMRKV</mark> G	gs <mark>cidtnw</mark> sa	 PNTGECD <mark>VSY</mark> TTTT EEE	KVDYLDSNN	NV <mark>IVHSKVF</mark>	HNKEL <mark>KTES</mark>	CD <mark>TKVVSN</mark> TI	L <mark>SVRVTVVS</mark> NDTRV	VEPTNESI	
A A <mark>IISIK</mark>	SLKRKMDO	B  <mark>CIFTEWTP</mark> KH	VDKCS <mark>IYYEV</mark>	EYRDRRND	VLWRENT TTI	N <mark>EAIFC</mark> SNV	VSASKVN <mark>SVW</mark>	F    <mark>IRGVE</mark> LNEQN <mark>VY</mark> GI SEEEE TTTTEE	DW <mark>YI HA</mark> L	
FKPDQV	QILSSVVF	B  K <mark>GR<mark>CVKTNWK</mark></mark>	llntgkcnv <mark>t</mark>	YYIEYK VN:	BEN <mark>KAIHHA</mark> S	SVE <mark>GEN</mark> SYE	EYCMEN <mark>ILDVI</mark>	F  <mark>Khtgtvymraaf</mark> gi Hheeeeeeeeti	IKGLW <mark>NNLR</mark> M	
SLPKPV	ENFKYYPM	b <mark>1</mark> ggn <mark>Cfkftw</mark>	LAQNTGLCK <mark>T</mark>	HFELQLLL	KNNK <mark>VKRQS</mark>	ISPMTTG <mark>T</mark> I	FIHCEVD <mark>SAS</mark> I	F LKK <mark>IYKAKIRSIYO</mark> HHHEEEEEEEEE	DGA <mark>TRLE</mark> GRW <mark>SL</mark>	LE L
FTPENV	KVTEAYLF	b  Eq <mark>cvtvrft</mark>	TLDVGTCK <mark>LS</mark>	YEFNYF DDI	RAQ <mark>LVGSSA</mark>	DKNTNTV	200 GITAS TVI	-F   G- <mark>Karars</mark> SD <mark>SV</mark> GQW <mark>S</mark> EEEEEETTEE E	SAYHI	
	IFN <mark>ATFYY</mark>	-B  (NSS <mark>CVYGTW</mark> )	NEENTGSCHL	NYHIQY DDI	NDA <mark>IHLT</mark> TK!	P <mark>EYTRC</mark> GL7	PNLK <mark>FVQMWA</mark>	-   G  <mark>Ay</mark> ng <mark>rv</mark> grk <mark>svys</mark> i Lettee eeee	:	
	R <mark>FSVNVK</mark> E	-B  EN <mark>CIFLTW</mark> KQ	PRTGLCAV <mark>SY</mark>	SVTLYGENI	DVLVYTHFN	SQVKESF	KYC <sup>SQLLRTII</sup>	F  NDIK <mark>TVGLQAVY</mark> KI 3G EEEEEEETI	NFGKVNRRHV	
SIPQKI	ENFQYAII	B  DT <mark>CFVLSWS</mark>	RQYTGNCIV <mark>N</mark>	HEIQYITGH	KN <mark>TTMNIDI</mark>	VTSNTNKLS	<mark>SYC</mark> APLPEDI	F  XK <mark>IKVIKIRSIY</mark> EF 3GEEEEEEEEETI	K <mark>RR</mark> GEW <mark>SSVN</mark> I	

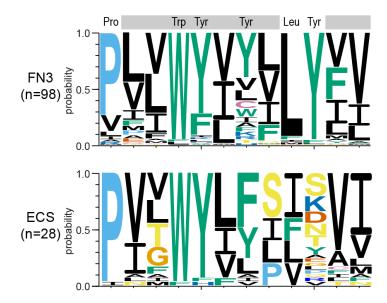
### Fig. S20. STRIDE secondary structure predictions for Alr domains 2 and 3

For each domain, the top line shows beta-strands labeled according to their position in the primary amino acid sequence. The middle line shows the sequence of the domain. The bottom line shows the STRIDE secondary structure predicted from the Colabfold model. (H = alpha helix, G = 3-10 helix, I = PI-helix, E = beta-strand extended conformation, B = isolated bridge, T = turn.)



# Fig. S21. Multiple sequence alignment of Fn3 domains and the Alr ECS fold.

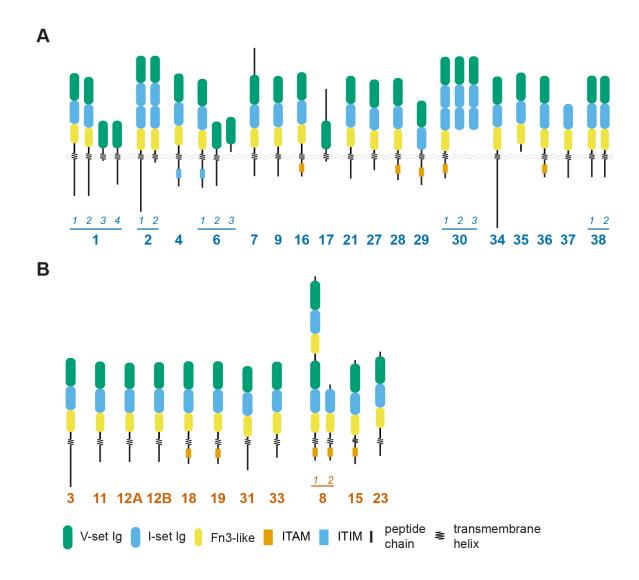
Alr ECS sequences were aligned to Fn3 domains from pfam. Residues in the alignment are highlighted by sequence conservation and chemical property with CLUSTALX colors as implemented in Jalview. The positions of residues typically conserved across Fn3 domains are shown above and below the alignment. The position of invariant cysteine residues is shown in red-orange lettering.



#### topohydrophobic position

Positive Charge (Arg, Lys, His) Negative Charge (Asp, Glu) Hydrophobic (Ala, Val, Iso, Leu, Met) Aromatic + Hydrophobic (Trp, Phe, Tyr) Polar (Ser, Thr, Asn, Gln) Cysteine Glycine Proline

**Fig. S22.** Sequence logo showing residues at conserved positions within Fn3 domains (top) and the ECS fold (bottom).



# Fig. S23. Predicted domains and motifs in *bona fide* and putative *Alr* gene products.

- (A) Bona fide Alr proteins
- (B) Putative Alr proteins

Genome			BAC contigs		
Assembly ID (length)	start	stop	ID (length)	start	stop
utg0000000001 (5,049,836 bp)	687,083	1	bc194 (1,225,536 bp)	1	684,144
utg000000021 (2,644,760 bp)	174	442,144	bc194 (1,225,536 bp)	783,992	1,225,536
utg0000000021 (2,644,760 bp)	1,005,148	1,590,913	bc18 (586,384 bp)	1	586,384
utg0000000121 (601,649 bp)	386,059	170,371	bc28 (214,692 bp)	1	214,692
utg0000000688 (716,359 bp)	88,750	244	bc050N15 (147,919 bp)	1	88,530
utg0000000026 (2,721,327 bp)	2,719,577	2,666,436	bc050N15 (147,919 bp)	94,782	147,919
utg0000000026 (2,721,327 bp)	2,661,684	2,138,742	bc174 (522,055 bp)	1	522,055
utg000000026 (2,721,327 bp)	1,818,447	1,610,954	bc29 (207,512 bp)	1	207,512

 Table S1. Overlap coordinates of genomic contigs and BAC contigs used to create reference ARC-F sequence.

	Gene model name	Expression	Reason for classifying as a pseudogene
-	Alr2p2.1	yes	Partial duplication of exons 1-4 of Alr2. Frame-shift in exon 3 leading to premature stop codons in exon 3.
	Alr2p2.2	yes	Partial duplication of exons 1-5 of Alr2. Frame-shift in exon 3 leading to premature stop codons in exon 3.
	Alr05p	yes	No evidence of splicing between exons 2-3.
	Alr10	yes	Improper splicing of exon 4 to downstream exons introduces stop codon in transcript.
	Alr12C	no	Stop codon in exon 2.
	Alr13	no	Stop codon in exon 1.
	Alr14p	no	No evidence of expression or exon encoding signal peptide.
	Alr20p	Yes	No evidence of expression or splicing between exons 1 and 2. No evidence of exon encoding a signal peptide.
	Alr22p	no	No evidence of expression or exon encoding signal peptide.
	Alr24p	yes	A few reads map to exons 2-3. No evidence of exon encoding a signal peptide.
	Alr25p	yes	A few reads map to exons 2-5. No evidence of exon encoding a signal peptide.
	Alr26p	yes	No open reading frame. No evidence of exon encoding a signal peptide. No splicing between exon 2-3.
	Alr32p	no	Only three exons, which have sequence similarity to Alr31, but are not expressed.

# Table S2. Gene models classified as Alr pseudogenes

Protein <sup>a</sup>	Domain		HMMER search of	pfam	HHpred search of SCOPe					
		Accession	description	e-value <sup>b</sup>	Accession	SCOPe Family	<i>probability</i> <sup>c</sup>			
Alr1	D1		•		d5l21b	b.1.1.1: V set domains	96.4			
Alr2	D1	PF07686.17	V-set	0.0012	d5e56a	b.1.1.1: V set domains	95.8			
Alr3	D1									
Alr4	D1				d5my6b1	b.1.1.1: V set domains	96.5			
Alr6	D1				d5my6b1	b.1.1.1: V set domains	96.5			
Alr7	D1				d5l21b	b.1.1.1: V set domains	96.0			
Alr8	Dla <sup>d</sup>				d5my6b1	b.1.1.1: V set domains	95.7			
Alr8	D1b <sup>e</sup>				d2esve1	b.1.1.1: V set domains	96.0			
Alr9	D1				d5my6b1	b.1.1.1: V set domains	96.1			
Alr11	D1				d5my6b1	b.1.1.1: V set domains	96.7			
Alr12A	D1				d4n8pa1	b.1.1.1: V set domains	96.6			
Alr12B	D1				d5my6b1	b.1.1.1: V set domains	96.9			
Alr15	D1									
Alr16	D1				d5my6b1	b.1.1.1: V set domains	96.9			
Alr17	D1				d5my6b1	b.1.1.1: V set domains	97.1			
Alr18	D1				d5my6b1	b.1.1.1: V set domains	97.0			
Alr19	D1				d5my6b1	b.1.1.1: V set domains	96.4			
Alr21	D1	PF07686.17	V-set	0.0012	d5my6b1	b.1.1.1: V set domains	96.8			
Alr23	D1				d5my6b1	b.1.1.1: V set domains	96.9			
Alr27	D1									
Alr28	D1	PF07686.17	V-set	1.40E-04	d5o04f1	b.1.1.1: V set domains	95.3			
Alr29	D1	PF07686.17	V-set	8.10E-05	d1yjdc1	b.1.1.1: V set domains	95.5			
Alr30	D1	PF07686.17	V-set	0.0031	d5e56a	b.1.1.1: V set domains	95.0			
Alr31	D1	PF17711.1	DUF5556	0.0097						
Alr33	D1									
Alr34	D1				d1c5db1	b.1.1.1: V set domains	88.8			
Alr35	D1	PF07686.17	V-set	1.00E-04	d5o04f1	b.1.1.1: V set domains	93.9			
Alr36	D1				d5my6b1	b.1.1.1: V set domains	93.4			
Alr38	D1				d5my6b1	b.1.1.1: V set domains	69.5			

# Table S3. Sequence homology of Domain 1

<sup>a</sup> proteins encoded by *bona fide* genes in blue, putative genes in red <sup>b</sup> significance cutoff = 0.01

<sup>d</sup> this is the membrane-distal domain with homology to other domain 1 sequences <sup>e</sup> this is the membrane-proximal domain with homology to other domain 1 sequences

<sup>c</sup> probability of homology; values <50% not shown; values >95% shaded in green

<b>Protein</b> <sup>a</sup>	Domain	Colabfold		DALI Top	Structural	Alignmen	t	PDBeFold					
		plDDT score <sup>b</sup>	PDB accession	Z-score <sup>c</sup>	RMSD	% ID	Domain Type <sup>d</sup>	PDB accession	Q-score <sup>e</sup>	RMSD	% ID	Domain Type <sup>d</sup>	
Alr1	D1	97.4	<u>7kqyE</u>	15.2	1.9	10	V-set	5uoeN	0.6356	1.589	9.9	V-set	
Alr2	D1	90.5	<u>30aiA</u>	15.1	2.1	20	V-set	<u>3m45C</u>	0.6243	1.663	14.7	V-set	
Alr3	D1	95.2	<u>3udwD</u>	14.8	1.9	14	V-set	<u>3udwD</u>	0.6095	1.565	14.6	V-set	
Alr4	D1	92.7	<u>5imkA</u>	14.4	1.8	16	V-set	<u>2iceT</u>	0.6337	1.708	16.5	V-set	
Alr6	D1	97.0	<u>603bE</u>	15.7	1.5	14	V-set	<u>5immA</u>	0.6647	1.695	13.6	V-set	
Alr7	D1	92.5	<u>2iceS</u>	14.9	2.1	11	V-set	<u>1neuA</u>	0.554	1.923	15.1	V-set	
Alr8	$D1a^{f}$	94.7	<u>2iceT</u>	15.2	2.1	13	V-set	<u>5immA</u>	0.6055	1.874	13.5	V-set	
Alr8	D1b <sup>g</sup>	89.3	<u>5imkA</u>	13.5	2.1	11	V-set	<u>5immA</u>	0.5658	1.875	12.3	V-set	
Alr9	D1	96.9	<u>603bE</u>	14.7	1.5	11	V-set	<u>6krzG</u>	0.6162	1.581	13.3	V-set	
Alr11	D1	96.7	<u>603bE</u>	15.7	1.6	14	V-set	2iceS	0.663	1.634	16.7	V-set	
Alr12A	D1	95.4	<u>5imkA</u>	14.7	1.8	15	V-set	2pndA	0.6482	1.708	15.1	V-set	
Alr12B	D1	92.8	<u>5imkA</u>	14.9	2	16	V-set	<u>2iceT</u>	0.6412	1.748	15.9	V-set	
Alr15	D1	85.6	<u>6bj2D</u>	14	2.1	15	V-set	<u>1u3hE</u>	0.5279	1.952	14.0	V-set	
Alr16	D1	95.7	<u>2iceT</u>	14.6	2.1	10	V-set	<u>2iceT</u>	0.6332	1.802	10.9	V-set	
Alr17	D1	95.5	<u>2iceS</u>	15	2.1	11	V-set	<u>2iceT</u>	0.6197	1.879	11.7	V-set	
Alr18	D1	95.7	<u>3qi9D</u>	14.9	1.8	10	V-set	<u>60ppL</u>	0.612	1.515	16.2	V-set	
Alr19	D1	97.3	<u>1tvdB</u>	15.3	2.2	9	V-set	5uoeN	0.6353	1.59	9.9	V-set	
Alr21	D1	95.0	<u>6j8gC</u>	14.4	2	13	V-set	<u>6j8hC</u>	0.6408	1.795	11.7	V-set	
Alr23	D1	95.9	2pndA	15.3	1.7	11	V-set	<u>5immA</u>	0.6338	1.668	12.2	V-set	
Alr27	D1	84.2	<u>2f53D</u>	14.1	2	9	V-set	<u>1u3hE</u>	0.5468	1.754	11.2	V-set	
Alr28	D1	80.6	<u>5m2wB</u>	14	2.1	9	V-set	<u>3ucrA</u>	0.5733	1.817	19.2	V-set	
Alr29	D1	88.7	<u>2iceT</u>	17.1	1.6	10	V-set	<u>2iceT</u>	0.7063	1.569	11.0	V-set	
Alr30	D1	92.8	<u>30aiA</u>	15.8	2	19	V-set	<u>6017L</u>	0.6295	1.419	19.4	V-set	
Alr31	D1	86.7	<u>2iceT</u>	14.2	2	13	V-set	2ptvA	0.6401	1.586	11.8	V-set	
Alr33	D1	84.3	<u>6dleB</u>	12.6	1.6	11	Ig domain	6arqA	0.5361	1.811	13.4	V-set	
Alr34	D1	95.7	<u>5imkA</u>	15	2.4	20	V-set	6vi4C	0.6021	2.113	20.5	V-set	
Alr35	D1	93.5	<u>1tvdA</u>	16.7	2.2	15	V-set	3b9kA	0.639	1.771	12.7	V-set	
Alr36	D1	92.9	<u>6fr6B</u>	15.3	1.9	13	V-set	<u>3udwA</u>	0.6006	1.884	14.9	V-set	
Alr38	D1	93.4	<u>5imlA</u>	15.8	2.2	16	V-set	2iccA	0.6086	1.944	16.1	V-set	

Table S4. Predicted structural homology for domain 1

v - set21ccA0.60861.94416.1V-sete Q-score = 1 are identical alignments; >0.5 are considered to have homologous structures

 Airso
 D1
 93.4
 Dimita
 13.8

 <sup>a</sup> bona fide genes in blue, putative genes in red

 <sup>b</sup> predicted local-distance difference test score; >90 considered highly accurate

 <sup>c</sup> Z-score between 8-20 indicates probable homology between query and hit

 <sup>d</sup> as annotated in the PDB (rcsb.org)

f this is the membrane-distal domain with homology to other domain 1 sequences in Alr8 <sup>g</sup> this is the membrane-proximal domain with homology to other domain 1 sequences in Alr8

### **Protein**<sup>a</sup> Domain **HMMER** search of pfam **HHpred search of SCOPe** SCOPe family *probability*<sup>c</sup> Accession description *e*-value<sup>b</sup> Accession Alr1 D2 PF07679.16 3.30E-05 b.1.1.4: I-set domains 97.6 I-set d1biha3 Alr2 D2 PF13927.6 Ig\_3 0.0002 d1biha3 b.1.1.4: I-set domains 99.0 D3 PF07679.16 0.0023 96.7 Alr2 I-set d1biha3 b.1.1.4: I-set domains Alr3 D2 d1biha3 b.1.1.4: I-set domains 96.6 Alr4 D2 d1biha3 98.7 b.1.1.4: I-set domains 97.0 Alr6 D2 PF07679.16 I-set 1.70E-05 d1biha3 b.1.1.4: I-set domains Alr7 D2 PF07679.16 d1x44a1 b.1.1.4: I-set domains 99.3 I-set 8.80E-05 D2a<sup>d</sup> PF07679.16 0.0024 d1biha3 97.4 Alr8 I-set b.1.1.4: I-set domains Alr8 PF07679.16 d1biha3 b.1.1.4: I-set domains 98.1 D2b<sup>e</sup> I-set 2.20E-07 Alr9 D2 d1biha3 97.4 b.1.1.4: I-set domains Alr11 D2 d1biha3 97.1 PF07679.16 I-set 0.0034 b.1.1.4: I-set domains Alr12A D2 PF13927.6 d1biha3 97.0 Ig 3 0.0028 b.1.1.4: I-set domains d1biha3 97.2 Alr12B D2 PF13927.6 Ig 3 0.0024 b.1.1.4: I-set domains Alr15 D2 d1biha3 b.1.1.4: I-set domains 84.9 97.4 Alr16 D2 PF07679.16 0.0021 d1biha3 b.1.1.4: I-set domains I-set 97.4 Alr18 D2 PF07679.16 I-set 3.50E-06 d1biha3 b.1.1.4: I-set domains Alr19 D2 PF07679.16 d1biha3 97.4 2.50E-06 b.1.1.4: I-set domains I-set Alr21 D2 PF07679.16 I-set 8.10E-05 d1biha3 b.1.1.4: I-set domains 97.6 Alr23 D2 PF07679.16 0.0005 d1biha3 97.5 I-set b.1.1.4: I-set domains Alr27 D2 Alr28 D2 PF07679.16 I-set 0.0076 d1vcaa2 b.1.1.4: I-set domains 97.4 Alr29 D2 85.2 d1ncua1 b.1.1.4: I-set domains D2 d1biha3 97.8 Alr30 PF13927.6 Ig 3 0.00027 b.1.1.4: I-set domains Alr30 D3 d1biha3 96.9 b.1.1.4: I-set domains Alr31 D2 d1ncua1 b.1.1.4: I-set domains 87.2 Alr33 D2 d1iray3 b.1.1.4: I-set domains 54.4 Alr34 D2 d1biha3 97.1 b.1.1.4: I-set domains D2 89.8 Alr35 PF07679.16 I-set 0.0066 d1koaa1 b.1.1.4: I-set domains Alr36 D2 d1biha3 97.4 b.1.1.4: I-set domains D2 d1biha3 Alr37 b.1.1.4: I-set domains 78.7 Alr38 D2 dliray3 b.1.1.4: I-set domains 54.7

### Table S5. Sequence homology for Domain 2 and 3

<sup>a</sup> proteins encoded by *bona fide* genes in blue, putative genes in orange

<sup>b</sup> significance cutoff = 0.01

<sup>d</sup> this is the membrane-distal domain with homology to other domain 1 sequences <sup>e</sup> this is the membrane-proximal domain with homology to other domain 1 sequences

<sup>c</sup> probability of homology; values <80% not shown; values >95% shaded in green

<b>Protein</b> <sup>a</sup>	Domain	Colabfold		DALI Top Structural Alignment			PDBeFold Top Structural Alignment					
		plDDT score <sup>b</sup>	PDB accession	<i>Z</i> -score <sup>c</sup>	RMSD	% ID	Domain Type <sup>d</sup>	PDB accession	Q-score <sup>e</sup>	RMSD	% ID	Domain Type <sup>d</sup>
Alr1	D2	95.1	<u>2rjmA</u>	12.9	1.7	19	I-set	<u>3qp3B</u>	0.5999	1.5	16	I-set
Alr2	D2	93.7	<u>1u2hA</u>	12.4	1.7	16	I-set	<u>1u2hA</u>	0.6237	1.6	16	I-set
Alr2	D3	90.9	<u>6efyA</u>	12.9	1.5	23	I-set	<u>3qp3C</u>	0.6663	1.3	15	I-set
Alr3	D2	88.2	<u>2rikA</u>	12.7	1.8	11	I-set	<u>6h4lA</u>	0.6153	1.7	13	I-set
Alr4	D2	94.4	<u>2j8hA</u>	13.2	1.4	18	I-set	<u>3pucA</u>	0.6466	1.3	15	I-set
Alr6	D2	93.9	<u>2rjmA</u>	12.2	1.7	15	I-set	<u>6h4lA</u>	0.6019	1.5	12	I-set
Alr7	D2	92.1	<u>2rjmA</u>	13.0	1.5	16	I-set	<u>4uowK</u>	0.6176	1.5	16	I-set
Alr8	D2a	88.0	<u>2rjmA</u>	13.4	1.8	21	I-set	<u>1u2hA</u>	0.6517	1.5	18	I-set
Alr8	D2b	92.5	<u>2rjmA</u>	13.6	1.4	16	I-set	<u>6h4lA</u>	0.6649	1.4	17	I-set
Alr9	D2	81.1	<u>4pgzA</u>	10.5	2.6	16	I-set	<u>3j9f8</u>	0.5011	2.3	12	I-set/C2-set
Alr11	D2	88.2	<u>2illA</u>	12.1	1.6	21	I-set	<u>1g1cB</u>	0.5566	1.7	19	I-set
Alr12A	D2	88.4	<u>3pucA</u>	12.9	1.6	10	I-set	<u>lg1cA</u>	0.5729	1.7	23	I-set
Alr12B	D2	85.7	<u>40f8B</u>	12.0	2.1	13	I-set/C2-set	<u>2wwmT</u>	0.5027	2.1	22	I-set
Alr15	D2	92.6	<u>40f8B</u>	10.8	2.1	12	I-set/C2-set	<u>3rghB</u>	0.5252	2.0	5	filamin
Alr16	D2	85.9	<u>4uow5</u>	11.5	2.1	19	I-set	<u>3j9f8</u>	0.4915	2.4	8	I-set/C2-set
Alr18	D2	89.4	<u>2rjmA</u>	13.1	1.7	24	I-set	<u>4uowG</u>	0.6208	1.6	22	I-set
Alr19	D2	92.6	<u>2rjmA</u>	12.1	1.8	23	I-set	<u>lg1cA</u>	0.588	1.5	23	I-set
Alr21	D2	87.5	<u>6efyA</u>	12.9	2.0	12	I-set	<u>6h4lA</u>	0.6162	1.5	8	I-set
Alr23	D2	92.2	<u>4pgzB</u>	11.6	2.3	14	I-set	<u>6h4lA</u>	0.5576	1.8	19	I-set
Alr27	D2	92.4	<u>3sbwC</u>	10.7	2.3	14	I-set/C2-set	<u>4uowB</u>	0.4968	1.8	14	I-set
Alr28	D2	91.8	<u>4pgzB</u>	13.0	1.9	16	I-set	<u>6h4lA</u>	0.6368	1.3	16	I-set
Alr29	D2	86.9	<u>40f8B</u>	10.6	2.2	15	I-set/C2-set	<u>2wwkT</u>	0.493	1.8	16	I-set
Alr30	D2	90.4	<u>1u2hA</u>	12.9	1.6	23	I-set	<u>1u2hA</u>	0.6853	1.4	24	I-set
Alr30	D3	92.1	<u>2fdbP</u>	12.3	1.8	12	I-set	<u>4uowE</u>	0.5936	1.7	12	I-set
Alr31	D2	84.0	<u>3dmkC</u>	11.4	2.4	13	I-set	<u>6h4lA</u>	0.5577	1.7	15	I-set
Alr33	D2	89.0	<u>6pv9A</u>	10.4	2.2	7	I-set/C2-set	<u>2kdgA</u>	0.5145	1.8	21	I-set
Alr34	D2	93.4	<u>2j8hA</u>	12.8	1.5	20	I-set	<u>3pucA</u>	0.6343	1.5	17	I-set
Alr35	D2	95.5	<u>3dmkC</u>	12.4	2.3	15	I-set	<u>6h4lA</u>	0.5833	1.9	16	I-set
Alr36	D2	90.0	2rikA	13.4	1.7	19	I-set	<u>6h4lA</u>	0.6571	1.4	13	I-set
Alr37	D2	92.6	<u>4uowR</u>	10.9	2.2	13	I-set	<u>4uowN</u>	0.5459	1.9	12	I-set
Alr38	D2	91.7	<u>2rikA</u>	12.5	1.7	16	I-set	<u>1u2hA</u>	0.6147	1.4	17	I-set

 Table S6. Predicted structural homology for domains 2 and 3

<sup>a</sup> *bona fide* genes in blue, putative genes in red <sup>b</sup> predicted local-distance difference test score; >90 considered highly accurate <sup>c</sup> Z-score between 8-20 indicates probable homology between query and hit <sup>d</sup> as annotated in the PDB (rcsb.org)

<sup>e</sup> Q-score = 1 are identical alignments; >0.5 are considered to have homologous structures <sup>f</sup> this is the membrane-distal domain with homology to other domain 1 sequences in Alr8 <sup>g</sup> this is the membrane-proximal domain with homology to other domain 1 sequences in Alr8

<b>Protein</b> <sup>a</sup>	Domain		HMMER search	of pfam	HHpred search of SCOPe					
		Accession	description	<i>e-value<sup>b</sup></i>	Accession SCOPe family		<i>Probability</i> <sup>c</sup>			
Alr1	ECS	Accession	description	e-value	d1j8ka	b.1.2.1: Fibronectin type III	88.1			
Alr2	ECS				d1fyhb1	b.1.2.1: Fibronectin type III	84.7			
Alr3	ECS				d3s9db1	b.1.2.1: Fibronectin type III	76.6			
Alr4	ECS					51				
Alr6	ECS				d1j8ka	b.1.2.1: Fibronectin type III	88.9			
Alr7	ECS				d1fnfa1	b.1.2.1: Fibronectin type III	79.7			
Alr8	ECSa <sup>e</sup>				d1j8ka	b.1.2.1: Fibronectin type III	85.1			
Alr8	$ECSb^{\mathrm{f}}$				5					
Alr9	ECS				d1j8ka	b.1.2.1: Fibronectin type III	81.6			
Alr11	ECS				d1j8ka	b.1.2.1: Fibronectin type III	82.4			
Alr12A	ECS				d1j8ka	b.1.2.1: Fibronectin type III	81.3			
Alr12B	ECS				d1j8ka	b.1.2.1: Fibronectin type III	80.2			
Alr15	ECS									
Alr16	ECS				d1j8ka	b.1.2.1: Fibronectin type III	85.2			
Alr18	ECS				d1j8ka	b.1.2.1: Fibronectin type III	87.8			
Alr19	ECS				d1j8ka	b.1.2.1: Fibronectin type III	87.8			
Alr21	ECS				d1j8ka	b.1.2.1: Fibronectin type III	82.0			
Alr23	ECS				d1fnfa1	b.1.2.1: Fibronectin type III	79.8			
Alr27	ECS									
Alr28	ECS				d1fyhb1	b.1.2.1: Fibronectin type III	75.6			
Alr29	ECS									
Alr30.1	ECS				d1fyhb1	b.1.2.1: Fibronectin type III	70.9			
Alr30.3	ECS				d1fyhb1	b.1.2.1: Fibronectin type III	82.4			
Alr31	ECS				d3d85d3	b.1.2.1: Fibronectin type III	56.8			
Alr33	ECS	PF07403.13	DUF1505	0.0013						
Alr34	ECS									
Alr35	ECS									
Alr36	ECS				d1fnfa1	b.1.2.1: Fibronectin type III	80.8			
Alr37	ECS				d1fnfa1	b.1.2.1: Fibronectin type III	74.3			
Alr38	ECS				d2gysa2	b.1.2.1: Fibronectin type III	60.5			

# Table S7. Sequence homology of the ECS fold

<sup>a</sup> proteins encoded by *bona fide* genes in blue, putative genes in red <sup>b</sup> significance cutoff = 0.01

<sup>d</sup> this is the membrane-distal domain with homology to other domain 1 sequences <sup>e</sup> this is the membrane-proximal domain with homology to other domain 1 sequences

<sup>c</sup> probability of homology; values <50% not shown; values >95% shaded in green

<b>Protein</b> <sup>a</sup>	Domain	Colabfold		DALI Top Structural Alignment			PDBeFold Top Structural Alignment					
		plDDT score <sup>b</sup>	PDB accession	Z-score <sup>c</sup>	RMSD	% ID	Domain Type <sup>d</sup>	PDB accession	Q-score <sup>e</sup>	RMSD	% ID	Domain Type <sup>d</sup>
Alr1	ECS	90.0	6h41A	11.4	1.8	13	Fn3	7jguA	0.5781	1.5	16	Fn3
Alr2	ECS	95.6	5fn8A	12.9	1.5	22	Fn3	5dc0A	0.6019	1.6	12	Fn3
Alr3	ECS	94.1	5fn6A	12.4	1.7	10	Fn3	<u>5n48D</u>	0.5855	1.8	8	Fn3
Alr4	ECS	95.4	7e9jB	12.7	1.8	15	Fn3	7jgtA	0.5270	2.0	13	Fn3
Alr6	ECS	94.0	7e9kD	13.2	1.4	16	Fn3	<u>1jrhI</u>	0.5330	1.6	5	Fn3
Alr7	ECS	90.0	5fn6A	12.2	1.7	9	Fn3	5n48D	0.5501	1.9	10	Fn3
Alr8 <sup>g</sup>	ECSa	91.1	7e9jB	13.0	1.5	14	Fn3	2rb8A	0.5490	2.2	10	Fn3
Alr8 <sup>h</sup>	ECSb	92.1	5fn6A	13.4	1.8	10	Fn3	2rb8A	0.5852	1.8	6	Fn3
Alr9	ECS	94.7	5fn8A	13.6	1.4	16	Fn3	7jguA	0.5956	1.6	16	Fn3
Alr11	ECS	95.1	5fn8A	10.5	2.6	18	Fn3	1tenA	0.5988	1.7	11	Fn3
Alr12A	ECS	95.3	5fn8A	12.1	1.6	14	Fn3	7jguA	0.5975	1.6	12	Fn3
Alr12B	ECS	94.2	5x83B	12.9	1.6	9	Fn3	7jguA	0.5947	1.6	11	Fn3
Alr15	ECS	93.9	2geeA	12.0	2.1	10	Fn3	3rzwA	0.6084	1.6	8	Fn3
Alr16	ECS	93.7	5fn6A	10.8	2.1	13	Fn3	1tenA	0.5925	1.8	6	Fn3
Alr18	ECS	93.1	6h41A	11.5	2.1	8	Fn3	7jguA	0.6015	1.6	12	Fn3
Alr19	ECS	94.5	5fn6A	13.1	1.7	14	Fn3	7jguA	0.5960	1.7	14	Fn3
Alr21	ECS	93.9	5fn6A	12.1	1.8	13	Fn3	5n48D	0.5829	1.9	10	Fn3
Alr23	ECS	84.7	6xfiA	12.9	2.0	15	Fn3	4wtwB	0.5724	1.6	18	Fn3
Alr27	ECS	94.3	2geeA	11.6	2.3	13	Fn3	50c7B	0.6465	1.7	15	Fn3
Alr28	ECS	92.4	3t1wA	10.7	2.3	11	Fn3	5n48D	0.6033	1.9	12	Fn3
Alr30.1	ECS	88.2	3t1wA	13.0	1.9	14	Fn3	4wtwA	0.5870	1.8	14	Fn3
Alr30.3	ECS	85.9	6mojB	10.6	2.2	12	Fn3	5n06A	0.4340	2.2	15	Fn3
Alr31	ECS	92.4	5fn8B	12.9	1.6	12	Fn3	5dc0A	0.5683	2.0	7	Fn3
Alr33	ECS	89.6	3t1wA	12.3	1.8	10	Fn3	2rb8A	0.6074	2.0	10	Fn3
Alr34	ECS	92.3	5n48D	11.4	2.4	7	Fn3	5dc0A	0.5811	2.0	7	Fn3
Alr35	ECS	96.5	5n48D	10.4	2.2	7	Fn3	5n48B	0.6154	1.9	7	Fn3
Alr36	ECS	93.1	5fn8B	12.8	1.5	8	Fn3	7jguA	0.5872	1.7	20	Fn3
Alr37	ECS	92.0	5n48D	12.4	2.3	9	Fn3	5n48D	0.5964	1.9	10	Fn3
Alr38	ECS	92.2	5n48D	13.4	1.7	11	Fn3	5n48D	0.6182	1.8	9	Fn3

Table S8. Predicted structural homology for the immunoglobulin-like fold in the ECS

<sup>e</sup> Q-score = 1 are identical alignments; >0.5 are considered to have homologous structures <sup>f</sup> this is the membrane-distal domain with homology to other domain 1 sequences in Alr8 <sup>g</sup> this is the membrane-proximal domain with homology to other domain 1 sequences in Alr8

<sup>a</sup> *bona fide* genes in blue, putative genes in red <sup>b</sup> predicted local-distance difference test score; >90 considered highly accurate <sup>c</sup> Z-score between 8-20 indicates probable homology between query and hit <sup>d</sup> as annotated in the PDB (rcsb.org)

<b>Protein</b> <sup>a</sup>	<b>Tandem Domains</b>	Colabfold
	I-set/FnIII-like	plDDT score <sup>b</sup>
Alr1	D2-ECS	92.2
Alr2	D3-ECS	92.5
Alr3	D2-ECS	89.8
Alr4	D2-ECS	95.2
Alr6	D2-ECS	93.6
Alr7	D2-ECS	92.7
Alr8	D2a-ECSa	91.4
Alr8	D2b-ECSb	92.9
Alr9	D2-ECS	92.0
Alr11	D2-ECS	93.7
Alr12A	D2-ECS	92.6
Alr12B	D2-ECS	93.2
Alr15	D2-ECS	76.7
Alr16	D2-ECS	91.9
Alr18	D2-ECS	92.3
Alr19	D2-ECS	94.7
Alr21	D2-ECS	92.5
Alr23	D2-ECS	85.0
Alr27	D2-ECS	91.7
Alr28	D2-ECS	90.0
Alr30	D3-ECS	87.6
Alr31	D2-ECS	74.2
Alr33	D2-ECS	87.3
Alr34	D2-ECS	88.2
Alr35	D2-ECS	95.7
Alr36	D2-ECS	93.9
Alr37	D2-ECS	86.4
Alr38	D2-ECS	88.0

Table S9. Structural predictions of tandem I-set and FnIII-like domains

<sup>a</sup> *bona fide* genes in blue, putative genes in red <sup>b</sup> predicted local-distance difference test score; values >90 are considered highly accurate

(Note: All Datasets are plain text files)

- Dataset S1. FASTA formatted sequence of the ARC-F reference. Two gaps of unknown physical size are denoted with N's.
- Dataset S2. GFF3-formatted annotations of *Alr* genes in the ARC-F reference sequence.
- Dataset S3. FASTA formatted sequence of contig utg7180000000456, which contains *Alr37*.
- Dataset S4. GFF3-formatted annotation of *Alr*37 on contig utg718000000456.
- Dataset S5. FASTA-formatted sequence of contig utg7180000000115, which contains *Alr38*.
- Dataset S6. GFF3-formatted annotation of *Alr38* on contig utg718000000115
- Dataset S7. FASTA-formatted cDNA sequences of bona fide genes in the *Alr* gene family.
- Dataset S8. FASTA-formatted amino acid sequences of Alr proteins encoded by bona fide genes.
- Dataset S9. FASTA-formatted cDNA sequences of putative genes in the *Alr* gene family.
- Dataset S10. FASTA-formatted amino acid sequences of Alr proteins encoded by putative genes.
- Dataset S11. FASTA-formatted MAFFT alignment of amino acid sequences for domain 1 from *bona fide* and putative *Alr* genes.
- Dataset S12. FASTA-formatted MAFFT alignment of amino acid sequences for domains 2 and 3 from *bona fide* and putative *Alr* genes.
- Dataset S13. FASTA-formatted MAFFT alignment of amino acid sequences for the ECS from *bona fide* and putative *Alr* genes.
- Dataset S14. FASTA-formatted amino acid sequences of the trimmed ECS used for structural predictions and alignment to fibronectin III domains.

## SI References

1. Cadavid LF, Powell AE, Nicotra ML, Moreno M, Buss LW. An invertebrate histocompatibility complex. Genetics. 2004;167: 357–365.

2. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33: 623–630.

3. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10: 563–569.

4. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9: e112963.

5. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5: R12.

6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

7. Noe L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research. 2005. pp. W540–W543. doi:10.1093/nar/gki478

8. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37: 907–915.

9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

10. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28: 511–515.

11. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol. 2019;15: e1006790.

12. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. In: RepeatMasker [Internet]. 2015 [cited 15 Jul 2020]. Available: http://www.repeatmasker.org

13. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12: 357–360.

14. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics. 2010;26: 1899–1900.

15. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2 a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25: 1189–1191. 16. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.

17. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: W256–W259.

18. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37: 420–423.

19. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305: 567–580.

20. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432.

21. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430: 2237–2243.

22. Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. bioRxiv. 2021. p. 2021.08.15.456425. doi:10.1101/2021.08.15.456425

23. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Research. 2004. pp. W500–W502. doi:10.1093/nar/gkh429

24. Holm L. Using Dali for Protein Structure Comparison. In: Gáspári Z, editor. Structural Bioinformatics: Methods and Protocols. New York, NY: Springer US; 2020. pp. 29–42.

25. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004;60: 2256–2268.

26. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.3. 2020.

27. Powell AE, Nicotra ML, Moreno MA, Lakkis FG, Dellaporta SL, Buss LW. Differential effect of allorecognition loci on phenotype in *Hydractinia symbiolongicarpus* (Cnidaria: Hydrozoa). Genetics. 2007;177: 2101–2107.

28. Nicotra ML, Powell AE, Rosengarten RD, Moreno M, Grimwood J, Lakkis FG, et al. A hypervariable invertebrate allodeterminant. Curr Biol. 2009;19: 583–589.

29. Rosa SF, Powell AE, Rosengarten RD, Nicotra ML, Moreno MA, Grimwood J, et al. *Hydractinia* allodeterminant *alr1* resides in an immunoglobulin superfamily-like gene complex. Curr Biol. 2010;20: 1122–1127.

30. Cannon JP, Haire RN, Litman GW. Identification of diversified genes that contain immunoglobulin-like variable regions in a protochordate. Nat Immunol. 2002;3: 1200–1207.