Title:

The *Hydractinia* Allorecognition Complex encodes a novel family of immunoglobulin superfamily proteins

Short Title:

The *Hydractinia* Allorecognition Complex

Authors:

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

Affiliations:

[1] Starzl Transplantation Institute, University of Pittsburgh, Pittsburgh, PA, USA
[2] Center for Evolutionary Biology and Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[3] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[4] NIH Intramural Sequencing Center, National Institutes of Health, Rockville, MD, USA
[5] Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL, USA
[6] Department of Biology, University of Florida, Gainesville, FL USA
[7] Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA

Corresponding Author: Matthew L. Nicotra matthew.nicotra@pitt.edu

ABSTRACT

Invertebrate histocompatibility—also known as allorecognition—has long interested marine ecologists, population geneticists, evolutionary biologists, and immunologists, but its genetic basis remains enigmatic in most species. Here, we report the nearly complete sequence of a histocompatibility complex from the colonial cnidarian, *Hydractinia symbiolongicarpus*. This sequence reveals that the two known *Hydractinia* allorecognition genes, *Allorecognition 1* (*Alr1*) and *Allorecognition 2* (*Alr2*) are part of a large family of immunoglobulin superfamily (IgSF) genes, several of which are candidates for new allodeterminants. These genes encode transmembrane proteins with domain architectures similar to cell-adhesion molecules and immune receptors. Several also contain cytoplasmic immunoreceptor tyrosine-based activation motifs (ITAMs) and immunoreceptor tyrosine-based inhibitory motifs (ITIMs). Our data, which include highly accurate protein structure predictions, reveal that these proteins have V-set and I-set domains with unusual sequence signatures. This suggests the last common ancestor of cnidarians and bilaterians had distinct V-set and I-set Ig domains.

INTRODUCTION

Allorecognition is the ability to distinguish self from non-self within the same species. Most encrusting colonial marine invertebrates, including sponges, corals, hydroids, bryozoans, and ascidians, are capable of allorecognition (Rosengarten and Nicotra, 2011). This ability enables colonies to compete with conspecifics for space but prevents them from competing with themselves as they grow on three-dimensional surfaces (Buss, 1990). Allorecognition also reduces the risk of stem cell parasitism, which can occur if unrelated colonies fuse and one colony's germline contributes disproportionately to the gametic output of the chimera (Laird et al., 2005).

Allorecognition has long attracted the attention of marine ecologists interested in spatial competition (Buss, 1990), population geneticists interested in the generation and maintenance of allelic diversity (Grosberg, 1988), and evolutionary biologists interested in units of selection and the origins of multicellularity (Buss, 1987; Grice and Degnan, 2015). In addition, ever since immunologists learned that corals and sea squirts exhibit allorecognition, they have wondered whether the genes that underlie this ability might be homologous to vertebrate histocompatibility genes (Burnet, 1971). If so, studying invertebrate allorecognition could help resolve the evolutionary history of immunity and perhaps lead to novel therapies in immunity and transplantation. Together, these interests have motivated the study of allorecognition genes in several species, including the poriferan *Amphimedon queenslandica* (Grice et al., 2017), the protochordate *Botryllus schlosseri* (De Tomaso et al., 2005; Nyholm et al., 2006; Voskoboynik et al., 2013), and the cnidarian *Hydractinia symbiolongicarpus* (Cadavid et al., 2004; Nicotra et al., 2009; Powell et al., 2007; Rosa et al., 2010).

In *Hydractinia*, allorecognition is controlled by the allorecognition complex (ARC), which encodes two linked genes called *Allorecognition 1* (*Alr1*) and *Allorecognition 2* (*Alr2*) (Cadavid et al., 2004; Powell et al., 2007). In laboratory strains, each gene has two alleles, and together they control allorecognition using a "missing-self" strategy. Colonies that share at least one allele at both genes recognize each other as self and fuse to create a larger colony (Fig. 1A). Colonies that do not share alleles at either locus recognize each other as non-self and fight by discharging harpoon-like organelles called nematocysts into their opponents (Fig. 1B). Colonies that share alleles only at one locus—either *Alr1* or *Alr2*—fuse but later separate.

Alr1 and *Alr2* encode transmembrane proteins with highly polymorphic extracellular domains (Nicotra et al., 2009; Rosa et al., 2010). In nature, there are tens to hundreds of alleles for each gene (Gloria-Soria et al., 2012; Rosa et al., 2010). *In vitro* studies have shown that the Alr1 protein is capable of *trans* (cell-to-cell) homophilic binding, which only occurs between allelic variants with similar extracellular sequences (Karadge et al., 2015). The same is true for Alr2 (Huene et al., 2021; Karadge et al., 2015). This variant-specific homophilic binding is hypothesized to be the mechanism of self/non-self discrimination *in vivo*.

The homology of *Alr1* and *Alr2* to other genes is unresolved. When they were originally

identified, it was not possible to identify orthologs for either gene outside of *Hydractinia*. However, BLAST searches did return statistically significant alignments between domains in the Alr1/2 extracellular regions and immunoglobulin (Ig) domains (Fig. 1C). These hits had low sequence identity, making it difficult to determine whether the Alr domains belonged to the immunoglobulin superfamily (IgSF).

It is also unclear how many *Alr* genes exist in *Hydractinia*. *Alr1* is flanked by several *Alr1*-like sequences (Rosa et al., 2010), and the genomic region immediately upstream *Alr2* contains two *Alr2* pseudogenes (Nicotra et al., 2009; Rosengarten et al., 2011). However, the full extent of this gene family is unknown because only a fraction of the ARC has been sequenced. Identifying additional *Alr* genes is of particular interest because the ARC most likely contains at least one additional allodeterminant. Evidence for this comes from the fact that *Alr1* and *Alr2* can fail to predict allorecognition responses in field-collected colonies (Nicotra et al., 2009; Rosa et al., 2010). The unidentified allodeterminant(s) likely reside in the ARC because genetic studies have shown that all *Hydractinia* allodeterminants, including dominant and codominant modifiers, are linked to *Alr1* and *Alr2* (Powell et al., 2011).

Here, we report the assembly and annotation of a nearly complete reference sequence of the ARC. We identify a family of 41 *Alr* genes, putative genes, and pseudogenes. A majority encode single-pass transmembrane proteins with extracellular V-set and I-set Ig domains and, unexpectedly, an additional fibronectin III (Fn3)-like fold. Several Alr proteins also have immunoreceptor tyrosine-based activation motifs (ITAMs) or immunoreceptor tyrosine-based inhibitory motifs (ITIMs) in their cytoplasmic tails. Together, these findings mark the discovery of a novel family of immunoglobulin superfamily (IgSF) proteins and provide a slate of candidates for additional allodeterminants in *Hydractinia*.

RESULTS

The ARC spans at least 11.8 Mb

In previous work, we generated a BAC library from a colony homozygous for the ARC-F haplotype (colony 833-8 in Fig. S1), and used it to perform chromosome walks starting from markers in the ARC linkage (Fig. 1D) (Nicotra et al., 2009; Powell et al., 2007; Rosa et al., 2010). The minimum tiling path of each walk was sequenced, resulting in six BAC contigs with a total length of 2.9 Mb (Fig. 1D and Fig. S2A). Here, we sequenced and assembled the genome of colony 236-21, an ARC-F homozygote and a descendant of colony 833-8 (Fig. S1). Genomic DNA was sequenced via PacBio long-read sequencing and polished with Illumina data to create a genome assembly that was 431 Mb long, with 5,697 contigs and an N50 of 224 Kb. We then aligned the original BAC contigs to this new assembly using Nucmer (Kurtz et al. 2004). We identified five genome contigs that overlapped the BAC contigs with >99% sequence identity (Fig. 1D, Fig. S2B, Table S1). The only major discrepancies between the BAC contigs and the genome contigs were in repeat regions. Therefore, we merged these sequences by filling the gaps between the BAC contigs with sequences from the genome assembly. The resulting ARC-F reference sequence spans 11.83 Mb and contains two gaps of unknown size (Fig. 1D, File

S1).

The ARC contains a large family of *Alr* genes and pseudogenes

We next annotated all *Alr*-like genes, guided by *ab initio* gene predictions, sequence similarity to *Alr1* and *Alr2*, and RNAseq data. Sequences without homology to *Alr1* or *Alr2* were not annotated. As new gene models were created, we used them in iterative TBLASTX searches to identify *Alr* genes that might not have been detected in earlier similarity searches. Finally, to identify *Alr* genes that exist outside the ARC-F reference sequence, we used TBLASTX to query the full genome assembly with the amino acid translation of each *Alr* gene model. All gene models were then numbered sequentially, with pseudogenes receiving a lowercase '*p*' at the end of their name (e.g., *Alr5p*). Alternative splice variants were indicated with a decimal number (e.g., *Alr1.2*). Gene models whose full-length predicted amino acid sequences had >80% sequence identity were given the same number followed by a letter (e.g., *Alr12A* and *Alr12B*).

In total, we created 41 gene models (Fig. 1E). All but two were located in the ARC reference sequence (File S2). More than half (27/41) were encoded in one of three *Alr* clusters, which we named A, B, and C (Fig. 1E). The remaining genes, *Alr37* and *Alr38*, were on contigs not contiguous with the reference sequence (Fig. 1E, Files S3-S6). The expression level of each gene model was estimated from our RNAseq data (Fig. 2A). Gene models with less than 1 FKPM were deemed unexpressed. To aid our analysis, we then classified each gene model as a *bona fide* gene, putative gene, or pseudogene.

A gene model was classified as a *bona fide* gene if it had one open reading frame (ORF) and each exon was expressed and properly spliced. Eighteen models fit this definition, including *Alr1* and *Alr2* (File S7 and S8). As shown in Fig. 2B, *Alr* genes generally encoded single-pass transmembrane proteins with 1-3 domains similar to the *Alr1* and *Alr2* Ig-like domains, an extracellular spacer (ECS), a transmembrane helix, and a cytoplasmic tail. Without exception, each Ig-like domain, ECS, and transmembrane helix was encoded by a single exon.

A gene model was classified as a putative gene if it had features that made us hesitant to call it a *bona fide* gene, but was not obviously a pseudogene. Eleven gene models fit this definition (File S9 and S10). Six had clear sequence similarity to *bona fide Alr* genes but were unexpressed (Fig. 2A, C). We did not call them pseudogenes because they could be expressed at developmental time points or in tissues not represented in our RNAseq dataset. Two gene models had ORFs that would encode a full Alr protein, but there was no evidence of splicing between exons 1, 2, and 3 (Fig. 2D). Three gene models did not encode a signal peptide (Fig. 2E).

A gene model was classified as a pseudogene if it was similar to a *bona fide* or putative *Alr* gene but was truncated by nonsense or frameshift mutations. Twelve gene models fit this definition (Table S2). Several pseudogenes were expressed at modest levels relative to other *Alr* genes (Fig. 2A).

We paid particular attention to the region directly upstream of *Alr2* because it contained two pseudogenes reported in previous publications (Nicotra et al., 2009; Rosengarten et al., 2011). The first, immediately upstream of *Alr2*, was named CDS6P by Nicotra *et al*. (2009) and *alr2P1* by Rosengarten *et al.* (2011). It was assumed to be a non-functional partial duplication of *Alr2*. Here, we identified additional exons encoding a transmembrane domain, cytoplasmic tail, and 3' UTR. We therefore classified this locus as a *bona fide* gene and named it *Alr30*. The second pseudogene was called CDS5P by Nicotra *et al*. (2009) and *alr2P2* by Rosengarten *et al.* (2011). Here, we also concluded the locus was a pseudogene. For consistency with the previous work, we have named this pseudogene *Alr2p2.*

Alternative splicing alters the domain architecture of several *Alr* gene products

Several *Alr* genes were alternatively spliced in ways that would change their gene product's domain architecture. *Alr1*, for example, had four splice variants. *Alr1.1* and *Alr1.2* were previously described (Rosa et al., 2010), but in *Alr1.3* and *Alr1.4*, exon 2 was spliced to new exons encoding alternative transmembrane domains and cytoplasmic tails (Fig. 2F). *Alr6* was also alternatively spliced in a similar manner (Fig. 2C). Notably, in *Alr6.3*, exon 2 was spliced to exon 11 and lacked a transmembrane helix, raising the possibility that its gene product is secreted. A similar splicing pattern, potentially leading to secreted gene products, was observed in Alr30 and Alr35 (Fig. 2A and S3A). At *Alr2*, transcripts lacking the 22-bp exon 7 were also detected, which would introduce a frameshift that truncated the cytoplasmic tail (Fig. 2A and S3B).

Sequences of Alr family members are highly diverse and undergo exon shuffling

The shared domain architecture of *Alr* proteins suggested a history of gene duplication. To investigate the evolutionary relationships between *Alr* family members, we attempted to create a single multiple sequence alignment for the gene products of all *Alr* genes and putative genes. However, their sequences were so divergent that it was impossible to obtain a high-quality alignment even after restricting ourselves to sequences of similar length. This drove us to assess overall sequence similarity within the family by performing all possible pairwise alignments. We found the average percent identity between any two amino acid sequences (excluding splice isoforms of the same gene) was 24.3% ± 8.6% (Fig. 3A). Only 2% of pairwise alignments had more than 50% identity. Thus, a substantial amount of sequence evolution has occurred since the origin of the *Alr* family.

Closer analysis of the pairwise alignments suggested a history of exon shuffling between *Alr* genes. Specifically, we found several cases in which two predicted proteins were very similar in some domains but very divergent in others (Fig. 3B). For example, Alr12A and Alr12B were >90% identical along their entire length except for domain 1, which was only 64% identical (Fig. 3B). This pattern of domain-level variation is consistent with previous studies indicating a history of exon shuffling between *Alr1* and nearby *Alr*-like sequences (Rosa et al., 2010) and between *Alr2* and nearby *Alr2* pseudogenes (Gloria-Soria et al., 2012). Our data indicate exon shuffling

could occur between several other *Alr* genes.

This pattern of exon shuffling led us to compare Alr proteins at the domain level. We subdivided each sequence into its constitutive extracellular domains, then produced multiple sequence alignments (File S11-S13) and neighbor-joining trees (Fig. 3C) for each domain type. This revealed a pattern in which domains were more similar if they were encoded close to each other in the genome (Fig. 3C). While this analysis has limited power to elucidate the history of the *Alr* gene family, it does suggest that the duplications within Cluster C occurred after it split from Clusters A/B.

The cytoplasmic tails of many Alr proteins contain ITAMs or ITIMs

Unlike their extracellular domains, the Alr cytoplasmic tails were too diverse to be included in a single alignment. Therefore, we used CD-HIT to cluster them at 20% sequence identity. This placed 16/32 into three groups, for which we created separate alignments (Fig. 4A, Fig. S4). Of the remaining 16 tails, three were <14 aa and the rest could not be grouped with other sequences. Thus the cytoplasmic tails of Alr proteins are more divergent than their extracellular domains.

The domain architecture of most Alr proteins suggested they might be receptors with intracellular signaling functions. To investigate this, we searched their cytoplasmic tails for signaling motifs. We found immunoreceptor tyrosine-based activation motifs (ITAMs) in the tails of six *bona fide* and eight putative Alr proteins (Fig. 4B-D). ITAMs, which have a consensus sequence of Yxx[I/L]$x_{(6-9)}$Yxx[L/I] (Murphy et al., 2008), are found in receptors that activate immune responses in vertebrates (Bezbradica and Medzhitov, 2012) and phagocytosis of damaged cells in *Drosophila* (Ziegenfuss et al., 2008).

Phosphorylated ITAMs are bound by the dual SH2 domains of a kinase called *Syk* in vertebrates and *Shark* in insects (Mócsai et al., 2010). Syk and Shark are related proteins and differ in that Shark has a set of ankyrin repeats between its two SH2 domains. To determine whether the *Hydractinia* genome encodes *Syk* or *Shark*-like kinases that might bind to these ITAMs, we performed a TBLASTN search of the complete genome assembly with the amino acid sequences of human *Syk* and *Drosophila Shark*, identifying *Hydractinia* homologs of each (Fig. 4E). This is consistent with previous work that has identified *Syk*-like and *Shark*-like kinases in *Hydra* (Chan et al., 1994; Steele et al., 1999).

A second motif in vertebrates called the immunoreceptor tyrosine-based inhibitory motif (ITIM) is found in receptors that counteract ITAM-mediated signaling and downregulate immune responses (Lanier, 2008). We found ITIMs, defined as [I/L/V/S]xYxx[I/V/L] (Barrow and Trowsdale, 2006), in two Alr tails, both from group 1 (Fig. 4F). In mammals, phosphorylated ITIMs are bound by the SH2 domains of SHP-1 and SHP-2, two phosphatases that dephosphorylate ITAM-bearing receptors, Syk-like proteins, and other components of activating pathways (Garg et al., 2020; Lorenz, 2009). ITIMs are also bound by the phosphoinositide phosphatases SHIP1 and SHIP2, which dampen immune cell activation via the PI3K pathway

(Pauls and Marshall, 2017; Thomas et al., 2017). We searched the *Hydractinia* genome and identified four homologs of SHP-1/2 and one SHIP1/2 homolog (Fig. 4G).

Together, these data show that many *Alr* genes have ITAMs and ITIMs, motifs that regulate the recognition of self and non-self in other animals. Moreover, the *Hydractinia* genome includes homologs of the enzymes that bind phosphorylated ITAMs or ITIMs and act as effectors of cellular activation or inhibition.

Domain 1 is a V-set Ig domain

Domain 1 of Alr1 and Alr2 was originally described as V-set-like (Nicotra et al., 2009; Rosa et al., 2010). To determine whether domain 1 was similar to V-set Ig domains in Alr3-Alr38, we first used HMMER to compare each sequence to Pfam, a database of hidden Markov models for protein families (El-Gebali et al., 2019). At an E-value cutoff of <0.01, only 6/29 sequences were similar to V-set domains (Table S3). We also used HHpred, a method that is more sensitive to remote homologies (Zimmermann et al., 2018) to search the Structural Classification of Proteins extended (SCOPe) database, which classifies protein domains according to structural and evolutionary relationships (Chandonia et al., 2019). Using this approach, we found that 19/29 sequences had a >95% probability of homology to V-set domains (Table S3).

Homologous proteins can evolve such that their primary sequences become highly divergent but their structures remain relatively unchanged (Yang and Honig, 2000). Therefore, we predicted the tertiary structure of each domain 1 with Colabfold (Mirdita et al., 2021) to further investigate its homology (File S14). Colabfold is a Google Colaboratory implementation of AlphaFold2 (Jumper et al., 2021), which is capable of producing structural predictions with sub-angstrom root mean square deviation from experimental structures (Tunyasuvunakool et al., 2021). Each residue in a model produced by Colabfold is assigned a predicted local distance difference test (pLDDT) score, which estimates how well the prediction would agree with an experimental structure. Residues with pLDDT > 90 are considered highly accurate and have their side chains oriented correctly 80% of the time (Jumper et al., 2021; Tunyasuvunakool et al., 2021). Residues with pLDDT > 70 generally have their backbones predicted correctly. For the Alr domain 1 sequences, Colabfold produced the structural predictions with average (model-wide) pLDDT scores ranging from 80.6 to 97.4 (Fig. 5A and Table S4).

Next, we performed structural alignments against the Protein Data Bank (rcsb.org; Berman et al., 2000) with Dali (Holm, 2020) and PDBeFOLD (Krissinel and Henrick, 2004). The top hits from both methods were to V-set Ig domains (e.g., Fig. 5B) (Table S4). Structural alignments produced by Dali are assigned a Z-score, which is used to estimate the likelihood that the two proteins are homologous. Z-scores between 8-20 indicate likely homologs (Holm, 2020). The domain 1 alignments had Z-scores ranging from 12.6-16.7, indicating probable homology with V-set domains (Fig. 5A), even though their overall sequence identities were 9-20% (Table S4). Thus, our analysis of the primary, secondary, and tertiary structures of domain 1 all indicated they belong to the V-set family of Ig domains.

V-set domains have nine β-strands named A, B, C, C', C", D, E, F, and G according to their position in the primary amino acid sequence. Strand C" is only found in V-set domains. Strand A is split into A and A'. These β-strands are arranged as a Greek key to form a β-sandwich, with one β-sheet consisting of strands A, B, E, and D, and the other consisting of strands A', G, F, C, C', and C" (Fig. 5C). This is often referred to as the V-frame (Harpaz and Chothia, 1994).

To determine whether domain 1 has the V-frame, we used STRIDE (Heinig and Frishman, 2004) to determine the secondary structure of our models, then assigned letters to the β-strands (Fig. S5). Twenty five models had all nine β-strands (e.g., Fig. 5D). Four models were with the missing strand being in either the A or A' position (Fig. S5). Notably, all models had the V-set specific C" strand (Fig. S5). In addition, 17/29 models included a 7 aa α-helix between the C" and D strands, which is not typically found in V-set domains.

These results led us to question why HMMER did not identify many Alr domains as V-set domains. Previous studies of V-set domains have identified a set of eight residues that are highly conserved, even across domains with as little as 20% sequence identity (Cannon et al., 2002; Harpaz and Chothia, 1994). According to the nomenclature of Cannon et al (2002), they are Gly^{16}, Cys^{23}, Trp^{41}, Arg^{75}, Leu^{89} (or other hydrophobic), Asp^{98}, Tyr^{102}, and Cys^{104}. To determine whether these residues are conserved in domain 1, we generated a multiple sequence alignment between them and 60 canonical V-set sequences from the Pfam V-set sequence profile (pf07686). We then identified the residues that corresponded to the eight V-set residues (Fig. S6). Our findings are summarized in Fig. 5F.

In V-set domains, Cys^{23}, Trp^{41}, and Cys^{104} form a nearly invariant structural motif called the 'pin' (Lesk and Chothia, 1982). The cysteines form a disulfide bridge between β-strands B and F, while the tryptophan packs against the bond to stabilize the hydrophobic core of the β-sandwich. All Alr domain 1 sequences, however, lacked these Cys residues, and only two had the Trp. Instead, Cys^{25} was replaced by bulky, aromatic amino acids (Trp, Phe, or Tyr). Cys^{104} and Trp^{41} were replaced by hydrophobic amino acids. Thus, in domain 1, the 'pin' is replaced by a set of bulky hydrophobic residues that might serve a similar function by stabilizing the core of the β-sandwich.

The fourth and fifth V-set residues, Arg^{75} and Asp^{98}, form a salt bridge between the CD and EF loops. The salt bridge is thought to stabilize the "bottom" of the domain and is found only in the V-set and I-set immunoglobulin domains (Cannon et al., 2002; Harpaz and Chothia, 1994). We found the salt bridge in all but three (26/29), although the negatively charged Asp was often replaced with similarly-charged Glu. Thus, the salt bridge, a hallmark of V-set and I-set domains, is also present in domain 1.

The sixth canonical residue, Tyr^{102}, forms the 'tyrosine corner', a structural motif located at the start of the F strand and found only in Greek key proteins (Hemmingsen et al., 1994). While Tyr^{102} is highly conserved in V-set Ig-like domains (97% in our seed alignment), it was found in only 7/29 Alr domains (Fig. 5E). Instead, 20/29 had Phe, with its aromatic ring

occupying the same location as that of Tyr[102]. Mutational studies have shown that a Tyr❼Phe mutation has no effect on the ability of V-set Ig domains to fold properly (Hamill et al., 2000). Thus, the residues at position 102 are consistent with domain 1 folding like a V-set domain.

The seventh canonical residue is Gly[16], which is part of a β-turn between strands A' and B. A β-turn is a series of four residues that reverses 180° on itself such that the distance between $C^{\alpha}(i)$ and $C^{\alpha}(i+3)$ is less than 7 Å (Chou, 2000). β-turns often feature a hydrogen bond between CO(*i*) and NH(*i*+3) — the carboxyl and amine groups on the first and fourth residues, respectively — but this is not a requirement (Richardson, 1981). We found this β-turn in all Alr domain 1 models (Fig. S7). Twenty-eight featured a hydrogen bond, defined by the criterion that O(*i*) is< 3.5 Å from N(*i*+3) (Richardson, 1981). However, position *i* + *2* was Gly in only eleven sequences. Thus, like V-set domains, domain 1 is predicted to have a β-turn between strands A' and B, but in most cases it does not involve a glycine.

The eighth canonical V-set residue is a hydrophobic amino acid, typically leucine, at position 89. This residue resides at the center of the hydrophobic core. In Alr domain 1, 21/29 sequences had a leucine residue at this position. The remaining six had other hydrophobic residues. Thus, this canonical residue is shared between V-set and most domain 1 sequences.

In summary, the Alr V-set domains share some, but not all of the sequence motifs commonly considered diagnostic of the V-set family. In the case of the pin motif, tyrosine corner, and β-turn, the Alr sequences differ in a way that likely preserves the structural motif. Thus domain 1 appears to be a V-set Ig domain with a novel sequence profile.

Domains 2 and 3 are I-set Ig domains

We next explored the relationship between Ig domains and the 29 domain 2 and 3 sequences encoded by *bona fide* and putative *Alr* genes. At an E-value cutoff of <0.01, HMMER identified 14 as I-set Ig-like domains (pf07679) and another four as Ig-like domains (pf13927) (Table S5). HHpred indicated 23/32 had a >95% probability of being homologous to the I-set family (Table S5).

To further explore their similarity to I-set Ig domains, we generated structural predictions for each domain, with average pLDDT scores ranging from 81.1 to 95.1 (Fig. 6A, Table S6, and File S14). I-set domains are similar to V-set domains except that they lack a C" strand, and the C' strand is often shorter (Fig. 6B). We found that 21/31 domain 2 and 3 models had an I-set topology (e.g., Fig. 6C, Fig. S8), and three more were only missing one of the two split A strands. Six others were missing the C' strand (e.g., Fig. 6D, Fig. S8). These six models were I-set-like in having a split A strand, but similar to C1-set Ig domains in having a D strand and lacking a C' strand (Fig. 6B, Fig. S8). The remaining domain, from Alr9, was predicted to have only 6 β-strands and lacked both C' and D strands. When we searched for structural homologs, the top hits from PDBeFOLD and Dali were to I-set domains, although some hits were annotated as both I-set and C2-set domains, and the top hits Alr15 were to a filamin repeat (Fig. 6A, Table S6). Filamin repeats are in the "Early" (E-set) superfamily of immunoglobulin-like β-

folds and are possibly related to the IgSF or fibronectin type III superfamilies (Chandonia et al., 2019; Fox et al., 2014).

We next investigated whether domains 2 and 3 had any of the conserved sequence motifs found in I-set domains. To do so, we aligned the Alr domains to 48 canonical I-set domains from the Pfam I-set sequence profile (pf07679) (Fig. S9). We then searched for the sequence motifs common to V-frame Ig-like domains (Cannon et al., 2002; Harpaz and Chothia, 1994). These results are summarized as a sequence logo in Fig. 6F. With respect to the pin motif (C-W-C), all Alr domain 2 sequences had the central tryptophan, but the two domain 3 sequences had methionine in its place. All domains lacked the paired cysteines. One was replaced by a hydrophobic residue, and the second was replaced by residues bearing no consistent physicochemical property (Fig. 6F, Fig. S9). The Alr domains also lacked the salt bridge and tyrosine corner (Fig. 6F). The β turn between β-strands A' and B was present in all but three structural models (Fig. S10). The last of the eight conserved residues, a hydrophobic residue (typically leucine) was present, although in many it was a tyrosine (Fig. 6F).

More recently, Wang (2013) defined the sequence signature of I-set domains via nine sequence motifs, denoted *i* through *ix*. The first four motifs include the C-W-C pin (in motifs *i*, *ii*, and *iv*), the tyrosine corner (part of motif *i*), and the tight turn in the A'B loop (motif *iii*) (Fig. S9). We searched for the remaining five motifs and found that they were present in a majority of Alr domain 2 and 3 sequences (Fig. S9 and Fig. S11).

Taken together, our data indicate most domain 2 and 3 sequences belong to the I-set family of Ig domains. Although these domains lack the disulfide bridge, salt bridge, and tyrosine corner, they have most other sequence motifs associated with I-set domains, including the conserved tryptophan. Most are predicted to have an I-set fold, although several appear to have lost the C' strand.

Part of the ECS adopts a fibronectin type III-like fold

When Alr1 and Alr2 were described previously, no domains were identified in the ECS (Nicotra et al., 2009; Rosa et al., 2010). Here, we expanded our analysis to include all 29 ECS regions encoded by *bona fide* and putative *Alr* genes. Although HMMER searches against Pfam only return two hits to domains of unknown function, HHpred indicated part of the ECS had a 60-88% probability of homology to fibronectin type III (Fn3) domains (Table S7). To help us define this potential domain, we aligned the ECS sequences to the 98 Fn3 sequences in the seed alignment of the Pfam Fn3 profile (pf00041.23). We found that the N-terminal portion of the ECS aligned reasonably well to other Fn3 domains but the C-terminal portion did not (Fig. S12).

Because structure predictions are often better for single domains, we removed the C-terminal portion of the ECS sequences (Fig. S12, File S15) then predicted their structures with Colabfold. Twenty-six models had average pLDDT >90, with the remaining three models >80 (Fig. 7A, Table S8, and File S14). Next we investigated whether the Alr domains had a similar

topology to Fn3 domains. Fn3 domains are an Ig-like fold with seven β-strands arranged in the same topology as C2-set immunoglobulin domains (Halaby et al., 1999; Leahy et al., 1992) (Fig. 7B). All but one ECS model had seven β-strands (e.g., Fig. 7C), with the remaining model missing strand G (Fig. S13). In all models, the β-strands adopted a C2-set/Fn3 topology. We next searched the PDB for proteins with similar structures. All hits from Dali and PDBeFOLD were to Fn3 domains (Fig. 7A and Table S8). Thus, the secondary and tertiary structures of this Ig-like fold are predicted to be most similar to Fn3 domains.

The primary amino acid sequences of Fn3 domains have six conserved amino acids (Halaby et al., 1999; Leahy et al., 1992). To determine whether the Alr Fn3-like domain had these residues, we aligned them to sequences from the Pfam Fn3 profile (pf00041.23). The ECS sequences only aligned well to Fn3 domains at their N-terminus (Fig. S14). With respect to the six conserved amino acids, ECS and Fn3 sequences shared a proline at the beginning of strand A, a tryptophan at the end of strand B, and a tyrosine at the beginning of strand C (Fig. 7E and Fig. S14). The fourth conserved residue in Fn3 domains, a tyrosine at the end of strand C, was present in 8/29 ECS sequences, and replaced by phenylalanine in 14/29 ECS sequences. However, unlike Fn3 domains, the ECS sequences were missing the leucine in the EF loop and the tyrosine residue that forms the tyrosine corner in strand F. Fn3 domains also have six additional 'topohydrophobic' positions (i.e., positions usually occupied by VILFMWY residues) (Halaby et al., 1999), which were also present more than 50% of the time in the ECS sequences. (Fig. 7E and Fig. S14).

Taken together, these data indicate that most Alr proteins have a Fn3-like fold between their I-set domain and transmembrane helix.

Alr proteins have six invariant cysteines likely to form disulfide bridges

While investigating protein alignments of the Alr domains, we identified six cysteine residues that were conserved across all sequences. These residues are not typically found in immunoglobulin or Fn3 folds. Three were located in the Ig domain immediately preceding the Fn3-like domain, and three were located within the Fn3-like domain itself (Fig. 8A,B). Within the Ig domain, the first two cysteines were located in β-strands A and B, and were predicted to form a disulfide-bridge in 28/31 models. Within the Fn3-like domain, the first and third cysteines, located in β-strands B and E, were predicted to form a disulfide bridge in 28/31 models. Thus, an intra-domain disulfide bridge appears to stabilize the fold of most Alr I-set and Fn3-like folds.

The two remaining invariant cysteines were located in the EF loop of the Ig domain and the BC loop of the Fn3-like domain. These domains appear in tandem in 28 of the Alr proteins, raising the possibility that they are stabilized by a disulfide-bridge. To test this, we predicted structures for the tandem I-set Ig and Fn3-like domains from each Alr protein. All were predicted with high confidence (Table S9), and in all models the two domains were predicted to be linked by a disulfide bridge (e.g., Fig. 8C).

Alr orthologs may exist in other hydrozoans

To search for orthologs of the *Alr* gene family, we first queried *Hydra vulgaris* and *Clytia haemispherica*, the only hydrozoan species with publicly available proteomes. BLASTP searches against the predicted proteomes of each species returned hits with significant E-values, many of which were predicted to be single-pass transmembrane proteins (Table S10). Only two hits in *Clytia* were identified as having Ig domains by HMMER/Pfam. To broaden our search, we next queried the non-redundant protein database at NCBI via BLASTP, and excluded hydrozoans from the returned hits. For each Alr protein, the top scoring alignment was between its Ig domains and one or more Ig domains in the hit. Notably, proteins from the anthozoans *Nematostella* and *Acropora* were not among the top hits. These data suggest the *Hydractinia Alr* gene family may have evolved within hydrozoans.

DISCUSSION

We have shown that the *Hydractinia* ARC contains a family of genes homologous to *Alr1* and *Alr2*, which we have named the *Alr* gene family. Most encode proteins with a receptor-like topology and domain architecture (Fig. 8D). Despite this similarity, individual Alr proteins have low sequence identity when compared to each other, suggesting the gene family is old, has experienced high rates of molecular evolution, or both.

Our data indicate that the N-terminal domains of Alr proteins are Ig domains. Specifically, domain 1 is either part of the V-set family or represents a new family within the IgSF that is most closely related to V-set domains. Likewise, domains 2 and 3 are either part of the I-set family or represent a new family most closely related to I-set domains. We propose including the Alr domains within existing V-set and I-set families. One reason for doing so is that the sequence profiles often used to identify Ig domains may be biased against recognizing them in non-bilaterians. Indeed, profiles for V-set and I-set domains were originally defined using a handful of sequences, mostly from vertebrates and model organisms (Harpaz and Chothia, 1994; Williams and Barclay, 1988). Today V-set and I-set profiles in pfam contain thousands of sequences, but are dominated by sequences from chordates and, to a much lesser extent, arthropods (El-Gebali et al., 2019). Although there is considerable sequence diversity within these sequences, they still represent the descendents of only two evolutionary nodes within metazoans. A consequence, suggested by this study, is that many V-set and I-set Ig domains go undetected in non-bilaterians and other understudied taxa. As genomes continue to be sequenced, our ability to detect these domains has been enhanced by new methods for detecting remote homologs and producing highly accurate structural models (Jumper et al., 2021; Mirdita et al., 2021; Tunyasuvunakool et al., 2021). These new data will undoubtedly enhance our understanding of how Ig domains evolved and may force us to revisit how these families are defined.

Indeed, to our knowledge, this is the first time a V-set Ig domain has been identified outside of bilaterians. Our data therefore suggest the last common ancestor of *Hydractinia* and

bilaterians had distinct V-set and I-set Ig domains. These domains likely followed different evolutionary trajectories to arrive at their current sequences. Some Alr I-set domains appear to have subsequently lost the C' strand. Testing this hypothesis (or alternatives involving convergence) will require careful analysis of Ig domains from other metazoans, especially non-bilaterians.

We have also discovered that the region previously referred to as the ECS actually encodes a domain with an Fn3-like fold. However, the sequences of Alr and Fn3 domains differed substantially between strands D through G, and HHpred did not identify them as homologs with high confidence. It therefore remains unclear whether this Fn3-like domain belongs to the fibronectin III superfamily.

Many Alr Ig domains lack the disulfide bridge and tryptophan found at the core of most Ig domains. Although there are many examples of Ig domains lacking either the disulfide bridge or the tryptophan, we are unaware of any V-set or I-set Ig domains that lack both. The reason for this loss is unclear. Intriguingly, the Alr I-set domains, along with their neighboring Fn3-like domains, are predicted to have conserved disulfide bridges that link neighboring β-strands within one β-sheet. There is also a conserved disulfide bridge between neighboring I-set and Fn3-like domains, which may stabilize their orientation.

One obvious function for the newly described *Alr* genes is allorecognition. It is therefore significant that many *Alr* genes are located outside the genomic region originally mapped by Cadavid et al. (2004) and Powell et al. (2007). If these other *Alr* genes rendered homozygous in the inbred lines used for mapping, any effect they might have as an allodeterminant would have been missed. In nature, however, these genes might be as polymorphic as *Alr1* and *Alr2*. Such undiscovered allodeterminants could explain the appearance of unexpected allorecognition responses between outbred colonies (Nicotra et al., 2009; Rosa et al., 2010).

Alr6 is a good candidate for such an allodeterminant because it is very similar to *Alr1*. *Alr6* occupies a location in cluster A that is roughly syntenic with the location of *Alr1* in Cluster B. It has an *Alr1*-like intron/exon structure, and it encodes a cytoplasmic tail that is clearly homologous to that of *Alr1.* In addition, *Alr6* is alternatively spliced to generate isoforms that have a single Ig domain and an alternative cytoplasmic tail, a feature shared only with *Alr1*. We hypothesize this similarity could extend to *Alr6* functioning as a third allodeterminant in the ARC.

Alr30 could also be an allodeterminant. *Alr30* is immediately upstream of *Alr2* and was previously considered a pseudogene (Rosa et al., 2010; Rosengarten et al., 2011). Here, we show that *Alr30* encodes a transmembrane protein with an extracellular region with recognizable sequence similarity to *Alr2*, but a cytoplasmic tail that has no detectable similarity to other *Alr* genes. Since this gene resides within the genomic interval previously defined in Nicotra et al. (2009), it is formally possible that it, too, contributes to allorecognition phenotypes.

The domain architecture of most Alr proteins also suggests alternative functions in

extracellular protein-protein interactions. Tandem Ig domains are commonly found in cell adhesion molecules, proteins involved in cell-to-cell communication, and immune receptors (Barclay, 2003). An adhesive function would also be consistent with that already described for Alr1 and Alr2 (Karadge et al., 2015).

The ITAM motifs in some Alr cytoplasmic tails are also potentially significant. ITAM-mediated signaling activates inflammation and cellular immune responses in vertebrates (Mócsai et al., 2010). It also activates phagocytosis of unwanted or damaged cells in *Drosophila* (Evans et al., 2015; Ziegenfuss et al., 2008) and appears to promote immune responses to bacteria in oysters (Sun et al., 2020). An immune function for some ITAM-bearing Alr proteins therefore seems plausible.

Could ITAM-mediated signaling play a role in allorecognition responses? One possibility is that Alr proteins with ITAM motifs activate rejection responses when they bind non-polymorphic, *Hydractinia*-specific ligands on opposing tissues. This rejection response would then be inhibited if polymorphic allodeterminants bind a compatible ligand. At present, this model is only supported by two seemingly disparate observations. First, *Hydractinia* mount the most vigorous and sustained rejection responses against other *Hydractinia* (Lange et al., 1992). This indicates colonies can identify the type of tissue they encounter. Second, the initial stages of rejection and fusion are morphologically indistinguishable (Lange et al., 1989). In both responses, nematocytes migrate to the point of contact and arrange their nematocysts as batteries pointed at their opponent. In rejection, the batteries fire, but in a fusion, the nematocytes migrate away as the tissues merge. This suggests rejection could be the default allorecognition response in *Hydractinia*. In this model, Alr proteins with ITAMs would activate rejection, which would be inhibited later by homophilic binding between compatible allodeterminants. This model would be analogous to the balance of ITAM and ITIM-mediated signaling that determines whether natural killer (NK) cells become activated in the vertebrate immune system. If true, it could also indicate a deep evolutionary relationship between invertebrate and vertebrate self-recognition systems.

Our decision to classify some *Alr* sequences as putative genes relied heavily on whether the genes were expressed and correctly spliced. Two caveats are associated with our expression data. First, because the RNAseq experiment was primarily intended to guide our annotation, it did not include biological or technical replicates. The resulting expression levels should therefore be viewed as rough estimates. Second, the RNA used to generate these reads was extracted from a pool of feeding and reproductive polyps. Mat and stolon tissue — the normal sites of allorecognition responses — were not included because we and others have been unable to isolate high quality RNA from these tissues (unpublished data; Uri Frank, personal communication). It is also possible that some *Alr* genes are expressed at developmental time points not represented in our dataset. Therefore, we expect to update our classification with additional data in the future.

The sequences of the Alr genes themselves do not appear to be orthologous to other invertebrate allorecognition genes. Nonetheless, this new ARC sequence reinforces similarities

between *Hydractinia* and other species in which allorecognition is controlled by genomic clusters of related genes (Grice and Degnan, 2015). This clustering may enhance the efficiency and coordination of allorecognition gene expression. It may also facilitate the generation of sequence diversity via gene conversion or unequal crossing over. Moreover, these invertebrate allorecognition complexes are remarkably similar to the complexes that control self/non-self recognition in vertebrates, namely the major histocompatibility complex (MHC) (Kaufman, 2018), leukocyte receptor complex (LRC) (Trowsdale et al., 2015), and NK complex (NKC) (Kelley et al., 2005). Identifying evolutionary links between these systems may become possible in the future as we survey a broader swath of metazoan genomes and simultaneously deepen our molecular understanding of how invertebrate allorecognition works.

MATERIALS 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 aquaria filled with artificial seawater (Reef Crystals) as previously described (Cadavid et al., 2004). 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 (Berlin et al., 2015). 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 (Chin et al., 2013) followed by polishing with the Illumina short read data using the PilonGrid parallel wrapper (Walker et al., 2014).

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 (Kurtz et al., 2004). 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 (-l #). 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 (-l), 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 (Camacho et al., 2009) and dot plots generated in YASS (Noe and Kucherov, 2005). 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 µl 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 µg 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 (Kim et al., 2019). 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 (Li et al., 2009) before proceeding to quantitation. Using the reference annotations of the Alr genes, transcript abundance of the Alr genes was estimated with Cufflinks (Trapnell et al., 2010). Abundance estimates were corrected for multiple read mappings.

Annotation of *Alr* genes

Alr genes were annotated using Apollo (Dunn et al., 2019) 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 (https://www.repeatmasker.org) (Smit et al., 2015). 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 (Kim et al., 2015, 2019). 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 Alr proteins were performed using MAFFT (Katoh and Toh, 2010). 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. Clustering was performed with CD-HIT (Fu et al., 2012) 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 (Letunic and Bork, 2019), exported as scaled vector graphics files, and annotated in Adobe Illustrator.

Protein sequence analysis

Signal peptides were predicted with SignalP 5.0 (Armenteros et al., 2019). Transmembrane helices were predicted with TMHMM 2.0 (Krogh et al., 2001). Conserved protein domains were identified with the Pfam database (El-Gebali et al., 2019) using HMMER3 (http://hmmer.org/). For domain prediction by HHpred, sequences were submitted to the MPI Bioinformatics Toolkit (Zimmermann et al., 2018). The query MSA was generated via three iterations of HHblits against the Uniref30 database, with an e-value threshold of 1×10^{-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 (Mirdita et al., 2021). 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^{-3} 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 (Heinig and Frishman, 2004). The Alr structural model with the highest average pLDDT was then submitted to DALI (Holm, 2020) and PDBeFOLD (https://www.ebi.ac.uk/msd-srv/ssm/) (Krissinel and Henrick, 2004) 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 (Schrödinger, LLC, 2020).

REFERENCES

Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. *37*, 420–423.

Barclay, A.N. (2003). Membrane proteins with immunoglobulin-like domains--a master superfamily of interaction molecules. Semin. Immunol. *15*, 215–223.

Barrow, A.D., and Trowsdale, J. (2006). You say ITAM and I say ITIM, let's call the whole thing off: the ambiguity of immunoreceptor signalling. Eur. J. Immunol. *36*, 1646–1653.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M., and Phillippy, A.M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. *33*, 623–630.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bezbradica, J.S., and Medzhitov, R. (2012). Role of ITAM signaling module in signal integration. Curr. Opin. Immunol. *24*, 58–66.

Burnet, F.M. (1971). "Self-recognition" in colonial marine forms and flowering plants in relation to the evolution of immunity. Nature *232*, 230–235.

Buss, L.W. (1987). The Evolution of Individuality (Princeton, N.J.: Princeton University Press).

Buss, L.W. (1990). Competition within and between encrusting clonal invertebrates. Trends Ecol. Evol. *5*, 352–356.

Cadavid, L.F., Powell, A.E., Nicotra, M.L., Moreno, M., and Buss, L.W. (2004). An invertebrate histocompatibility complex. Genetics *167*, 357–365.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Cannon, J.P., Haire, R.N., and Litman, G.W. (2002). Identification of diversified genes that contain immunoglobulin-like variable regions in a protochordate. Nat. Immunol. *3*, 1200–1207.

Chan, T.A., Chu, C.A., Rauen, K.A., Kroiher, M., Tatarewicz, S.M., and Steele, R.E. (1994). Identification of a gene encoding a novel protein-tyrosine kinase containing SH2 domains and ankyrin-like repeats. Oncogene *9*, 1253–1259.

Chandonia, J.-M., Fox, N.K., and Brenner, S.E. (2019). SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. Nucleic Acids Res. *47*, D475–D481.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569.

Chou, K.C. (2000). Prediction of tight turns and their types in proteins. Anal. Biochem. *286*, 1–

16.

De Tomaso, A.W., Nyholm, S.V., Palmeri, K.J., Ishizuka, K.J., Ludington, W.B., Mitchel, K., and Weissman, I.L. (2005). Isolation and characterization of a protochordate histocompatibility locus. Nature *438*, 454–459.

Dunn, N.A., Unni, D.R., Diesh, C., Munoz-Torres, M., Harris, N.L., Yao, E., Rasche, H., Holmes, I.H., Elsik, C.G., and Lewis, S.E. (2019). Apollo: Democratizing genome annotation. PLoS Comput. Biol. *15*, e1006790.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432.

Evans, I.R., Rodrigues, F.S.L.M., Armitage, E.L., and Wood, W. (2015). Draper/CED-1 mediates an ancient damage response to control inflammatory blood cell migration in vivo. Curr. Biol. *25*, 1606–1612.

Fox, N.K., Brenner, S.E., and Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. *42*, D304-9.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Garg, M., Wahid, M., and Khan, F., Dr (2020). Regulation of peripheral and central immunity: Understanding the role of Src homology 2 domain-containing tyrosine phosphatases, SHP-1 & SHP-2. Immunobiology *225*, 151847.

Gloria-Soria, A., Moreno, M.A., Yund, P.O., Lakkis, F.G., Dellaporta, S.L., and Buss, L.W. (2012). Evolutionary genetics of the hydroid allodeterminant *alr2*. Mol. Biol. Evol. *29*, 3921–3932.

Grice, L.F., and Degnan, B.M. (2015). How to Build an Allorecognition System: A Guide for Prospective Multicellular Organisms. In Evolutionary Transitions to Multicellular Life: Principles and Mechanisms, I. Ruiz-Trillo, and A.M. Nedelcu, eds. (Dordrecht: Springer Netherlands), pp. 395–424.

Grice, L.F., Gauthier, M.E.A., Roper, K.E., Fernandez-Busquets, X., Degnan, S.M., and Degnan, B.M. (2017). Origin and Evolution of the Sponge Aggregation Factor Gene Family. Mol. Biol. Evol. *34*, 1083–1099.

Grosberg, R.K. (1988). The Evolution of Allorecognition Specificity in Clonal Invertebrates. Q. Rev. Biol. *63*, 377–412.

Halaby, D.M., Poupon, A., and Mornon, J. (1999). The immunoglobulin fold family: sequence analysis and 3D structure comparisons. Protein Eng. *12*, 563–571.

Hamill, S.J., Cota, E., Chothia, C., and Clarke, J. (2000). Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. J. Mol. Biol. *295*, 641–649.

Harpaz, Y., and Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell

adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J. Mol. Biol. *238*, 528–539.

Heinig, M., and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Research *32*, W500–W502.

Hemmingsen, J.M., Gernert, K.M., Richardson, J.S., and Richardson, D.C. (1994). The tyrosine corner: a feature of most Greek key beta-barrel proteins. Protein Sci. *3*, 1927–1937.

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

Huene, A.L., Chen, T., and Nicotra, M.L. (2021). New binding specificities evolve via point mutation in an invertebrate allorecognition gene. IScience *24*, 102811.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

Karadge, U.B., Gosto, M., and Nicotra, M.L. (2015). Allorecognition Proteins in an Invertebrate Exhibit Homophilic Interactions. Curr. Biol. *25*, 2845–2850.

Katoh, K., and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics *26*, 1899–1900.

Kaufman, J. (2018). Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates. Annu. Rev. Immunol. *36*, 383–409.

Kelley, J., Walter, L., and Trowsdale, J. (2005). Comparative genomics of natural killer cell receptor gene clusters. PLoS Genet. *1*, 129–139.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

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

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, R12.

Laird, D.J., De Tomaso, A.W., and Weissman, I.L. (2005). Stem cells are units of natural selection in a colonial ascidian. Cell *123*, 1351–1360.

Lange, R., Plickert, G., and Müller, W.A. (1989). Histoincompatibility in a low invertebrate,Hydractinia echinata: Analysis of the mechanism of rejection. J. Exp. Zool. *249*,

284–292.

Lange, R.G., Dick, M.H., and Müller, W.A. (1992). Specificity and early ontogeny of historecognition in the hydroidHydractinia. J. Exp. Zool. *262*, 307–316.

Lanier, L.L. (2008). Evolutionary struggles between NK cells and viruses. Nat. Rev. Immunol. *8*, 259–268.

Leahy, D.J., Hendrickson, W.A., Aukhil, I., and Erickson, H.P. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science *258*, 987–991.

Lesk, A.M., and Chothia, C. (1982). Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. *160*, 325–342.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lorenz, U. (2009). SHP-1 and SHP-2 in T cells: two phosphatases functioning at many levels. Immunol. Rev. *228*, 342–359.

Mirdita, M., Ovchinnikov, S., and Steinegger, M. (2021). ColabFold - Making protein folding accessible to all.

Mócsai, A., Ruland, J., and Tybulewicz, V.L.J. (2010). The SYK tyrosine kinase: a crucial player in diverse biological functions. Nat. Rev. Immunol. *10*, 387–402.

Murphy, K., Travers, P., and Walport, M. (2008). Janeway's immunobiology. Garland science. Taylor & Francis Group, LLC *15*, 655–708.

Nicotra, M.L., Powell, A.E., Rosengarten, R.D., Moreno, M., Grimwood, J., Lakkis, F.G., Dellaporta, S.L., and Buss, L.W. (2009). A hypervariable invertebrate allodeterminant. Curr. Biol. *19*, 583–589.

Noe, L., and Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research *33*, W540–W543.

Nyholm, S.V., Passegue, E., Ludington, W.B., Voskoboynik, A., Mitchel, K., Weissman, I.L., and De Tomaso, A.W. (2006). fester, A candidate allorecognition receptor from a primitive chordate. Immunity *25*, 163–173.

Pauls, S.D., and Marshall, A.J. (2017). Regulation of immune cell signaling by SHIP1: A phosphatase, scaffold protein, and potential therapeutic target. Eur. J. Immunol. *47*, 932–945.

Powell, A.E., Nicotra, M.L., Moreno, M.A., Lakkis, F.G., Dellaporta, S.L., and Buss, L.W. (2007). Differential effect of allorecognition loci on phenotype in *Hydractinia symbiolongicarpus* (Cnidaria: Hydrozoa). Genetics *177*, 2101–2107.

Powell, A.E., Moreno, M., Gloria-Soria, A., Lakkis, F.G., Dellaporta, S.L., and Buss, L.W. (2011).

Genetic background and allorecognition phenotype in *Hydractinia symbiolongicarpus*. G3 *1*, 499–504.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. Adv. Protein Chem. *34*, 167–339.

Rosa, S.F., Powell, A.E., Rosengarten, R.D., Nicotra, M.L., Moreno, M.A., Grimwood, J., Lakkis, F.G., Dellaporta, S.L., and Buss, L.W. (2010). *Hydractinia* allodeterminant *alr1* resides in an immunoglobulin superfamily-like gene complex. Curr. Biol. *20*, 1122–1127.

Rosengarten, R.D., and Nicotra, M.L. (2011). Model systems of invertebrate allorecognition. Curr. Biol. *21*, R82-92.

Rosengarten, R.D., Moreno, M.A., Lakkis, F.G., Buss, L.W., and Dellaporta, S.L. (2011). Genetic diversity of the allodeterminant *alr2* in *Hydractinia symbiolongicarpus*. Mol. Biol. Evol. *28*, 933–947.

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

Smit, A.F.A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0.

Steele, R.E., Stover, N.A., and Sakaguchi, M. (1999). Appearance and disappearance of Syk family protein-tyrosine kinase genes during metazoan evolution. Gene *239*, 91–97.

Sun, J., Wang, L., Yang, C., and Song, L. (2020). An Ancient BCR-like Signaling Promotes ICP Production and Hemocyte Phagocytosis in Oyster. IScience *23*, 100834.

Thomas, M.P., Erneux, C., and Potter, B.V.L. (2017). SHIP2: Structure, Function and Inhibition. Chembiochem *18*, 233–247.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trowsdale, J., Jones, D.C., Barrow, A.D., and Traherne, J.A. (2015). Surveillance of cell and tissue perturbation by receptors in the LRC. Immunol. Rev. *267*, 117–136.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596.

Voskoboynik, A., Newman, A.M., Corey, D.M., Sahoo, D., Pushkarev, D., Neff, N.F., Passarelli, B., Koh, W., Ishizuka, K.J., Palmeri, K.J., et al. (2013). Identification of a colonial chordate histocompatibility gene. Science *341*, 384–387.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

Wang, J.-H. (2013). The sequence signature of an Ig-fold. Protein Cell *4*, 569–572.

Williams, A.F., and Barclay, A.N. (1988). The Immunoglobulin Superfamily—Domains for Cell

Surface Recognition. Annu. Rev. Immunol. *6*, 381–405.

Yang, A.S., and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J. Mol. Biol. *301*, 679–689.

Ziegenfuss, J.S., Biswas, R., Avery, M.A., Hong, K., Sheehan, A.E., Yeung, Y.-G., Stanley, E.R., and Freeman, M.R. (2008). Draper-dependent glial phagocytic activity is mediated by Src and Syk family kinase signalling. Nature *453*, 935–939.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol. *430*, 2237–2243.

ACKNOWLEDGEMENTS

We thank Leo Buss for sharing colony 236-21 and for helpful discussion.

Funding:

National Science Foundation grant 1557339 awarded to (MLN)

National Science Foundation grant 1923259 awarded to (CES, MLN)

ALH and SMS were supported by National Institutes of Health grant T32 AI074490

National Institutes of Health, Intramural Research Program of the National Human Genome Research Institute, grant ZIA HG000140 (ADB)

National Institutes of Health, Intramural Research Program of the National Human Genome Institute, grant ZIA HG200398 to (AMP)

Author contributions:

Conceptualization: ALH, CES, JCM, ADB, MLN

Data Curation: ALH, SMS, AN, SK, CES, MLN

Formal Analysis: ALH, SMS, SK, CES, MLN

Funding acquisition: AMP, CES, ADB, MLN

Investigation: ALH, SMS, ZM, AN, SK, MHM, CES, MLN

Methodology: ALH, SMS, SK, AMP, ADB, MLN

Project Administration: CES, ADB, MLN

Resources: CES, ADB, MLN

Supervision: CES, ADB, MLN

Visualization: ALH, MLN

Writing – original draft: ALH, MLN

Writing – review and editing: ALH, SMS, SK, MHM, CES, ADB, MLN

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The raw PacBio and Illumina genomic reads, together with the RNAseq reads for colony 236-21 were deposited in the SRA under BioProject PRJNA802249. The 236-21 genome assembly is available in the NCBI Assembly database under Accession GCA_XXXXXX (and is available upon request to MLN). All other data are available in the main text or supplementary materials.

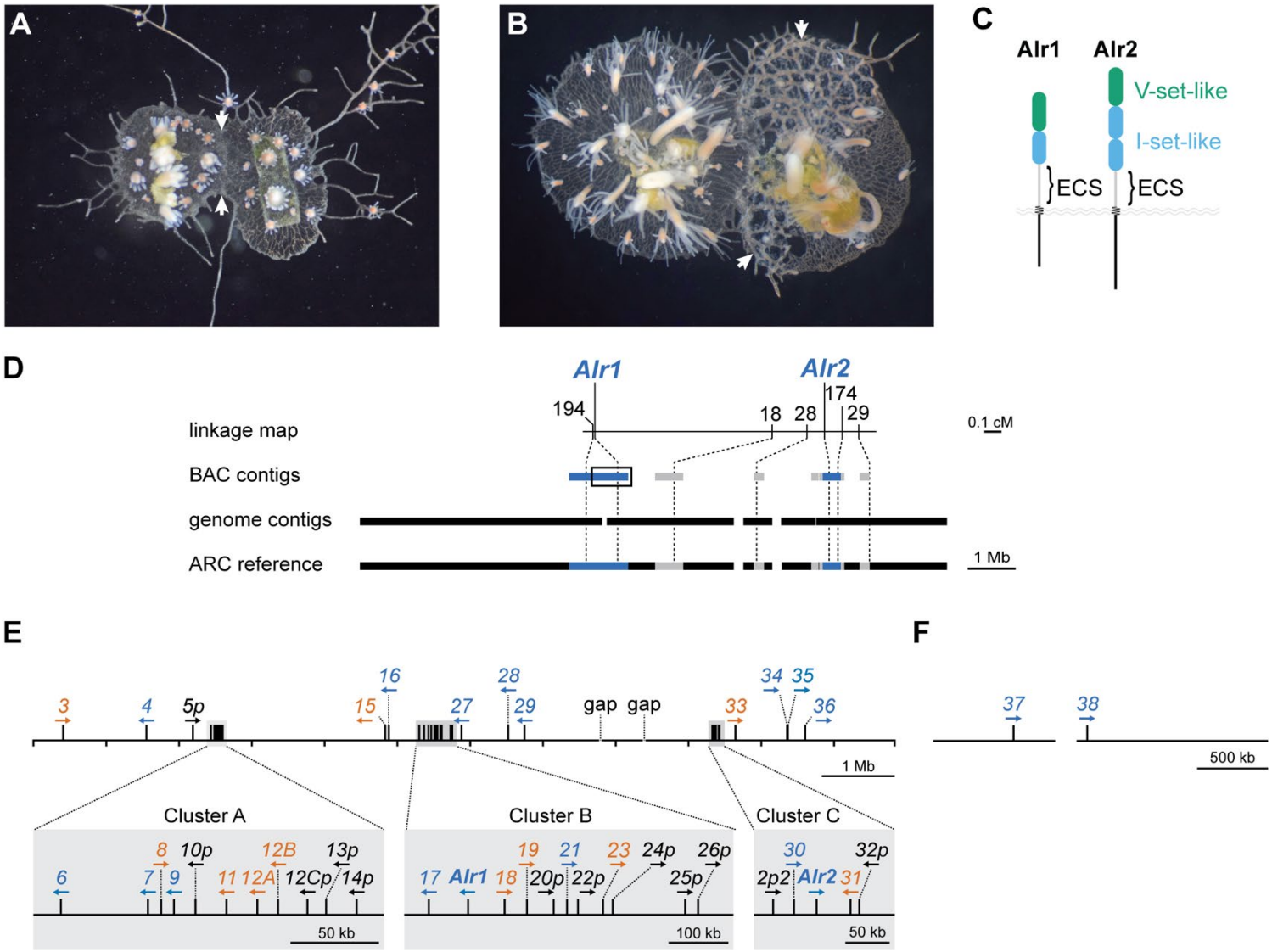

Fig. 1. Allorecognition in *Hydractinia* and the assembly of the *Alr* gene complex.

(A) Fusion between two compatible colonies. Arrows point to the region of fusion.

(B) Rejection between two incompatible colonies. The colony on the left has grown over the colony on the right. Arrows point to specialized structures called hyperplastic stolons, which are destroying the underlying tissue.

(C) Domain architecture of Alr1 and Alr2.

(D) Chromosome walks from the ARC linkage map (top) generated six BAC contigs (below; blue = previously published; gray = unpublished). These were aligned to contigs from the assembled genome of an animal homozygous across the ARC (black). The resulting 11.83 Mb reference sequence was constructed by concatenating the BAC and genome assemblies (bottom).

(E) Identity, location, and orientation of *Alr* family members within the ARC reference (blue, *bona fide* gene; orange, putative gene; black, pseudogene).

(F) Two *Alr* genes located in genome contigs that could not be physically linked to the ARC reference sequence.

Figure 2 (Legend on next page)

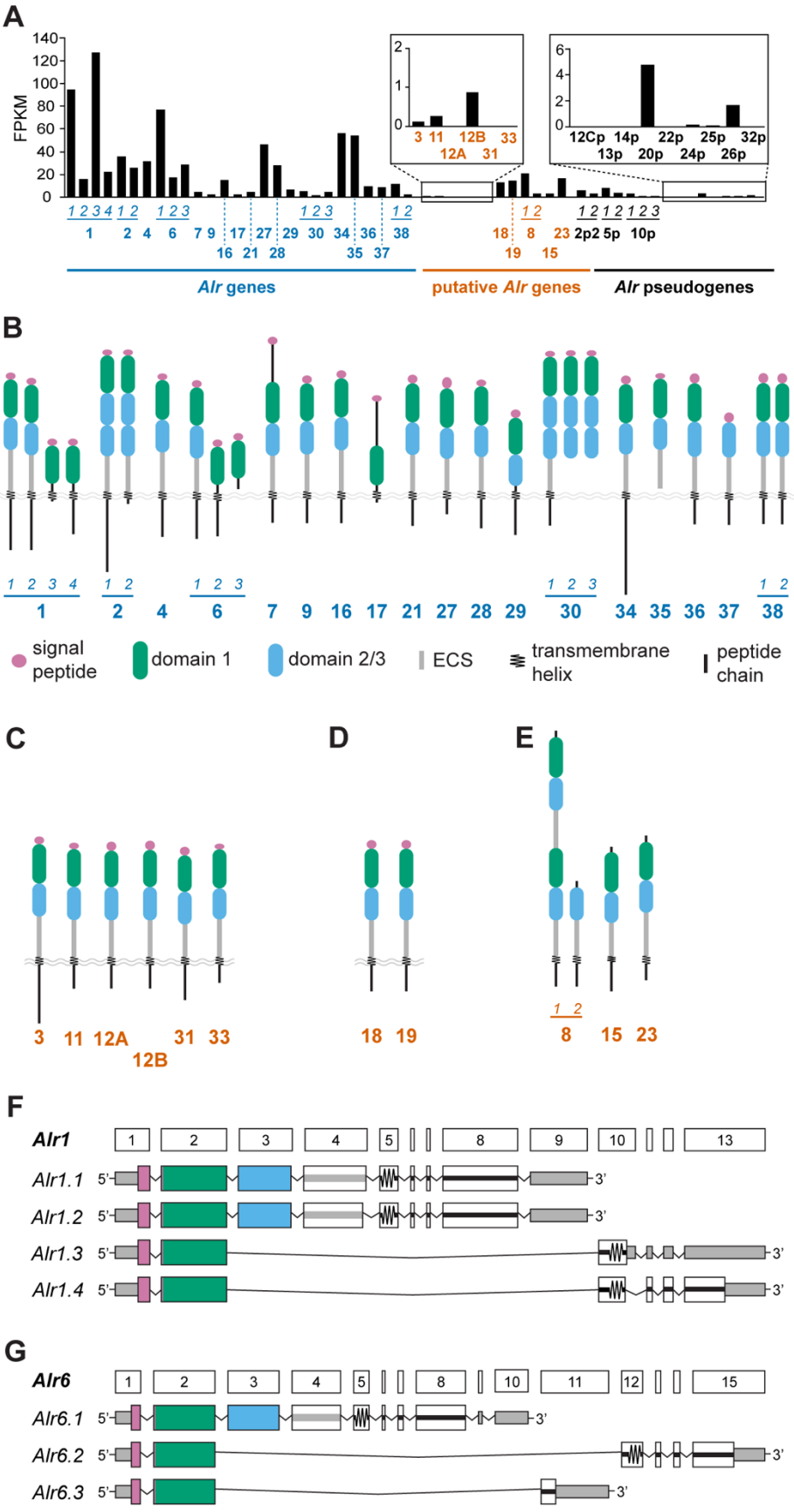

Fig. 2. Expression, domain architecture, and alternative splicing of *Alr* genes.

(A) Estimated expression level of each *bona fide* gene, putative gene, and pseudogene. FPKM, fragments per kilobase mapped. Genes are identified by bold numbers. Splice variants are indicated by horizontal lines and numbers in italics. The order of the putative genes follows their appearance in panels C-E.

(B) Domain architecture of *Alr* genes. Note that the signal peptide would be cleaved before surface expression, but is shown here to indicate its presence/absence in some gene models.

(C) Domain architecture of unexpressed putative genes.

(D) Domain architecture of partially expressed putative genes.

(E) Domain architecture of putative genes lacking a predicted signal peptide.

(F) Alternative splicing of *Alr1*.

(G) Alternative splicing of *Alr6.* In (F) and (G), exons are colored according to the type of domain/region they encode.

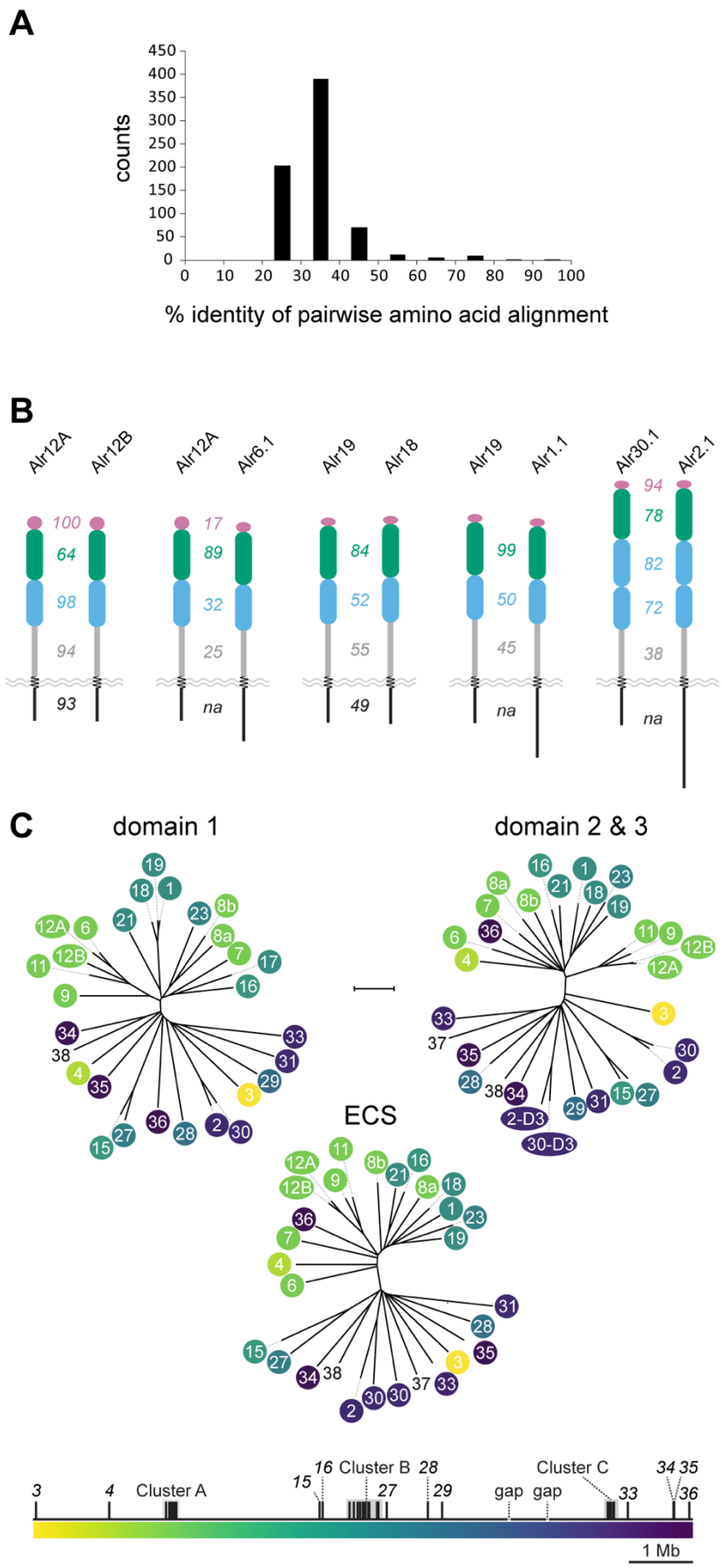

Fig. 3. Sequence similarity between Alr extracellular domains.

(A) Histogram of amino acid percent identities for pairwise alignments of *Alr* genes and putative genes.

(B) Evidence of exon shuffling between *Alr* genes. Numbers between protein pairs indicate sequence identity within that domain/region.

(C) Neighbor-joining trees of Alr extracellular domains. Leaves of each tree are color coded according to their genomic position. Domains from *Alr37* and *Alr38* are not color coded. Branch lengths calculated according to the BLOSUM26 matrix. Scale bar = 100 units.

Figure 4 (Legend on next page)

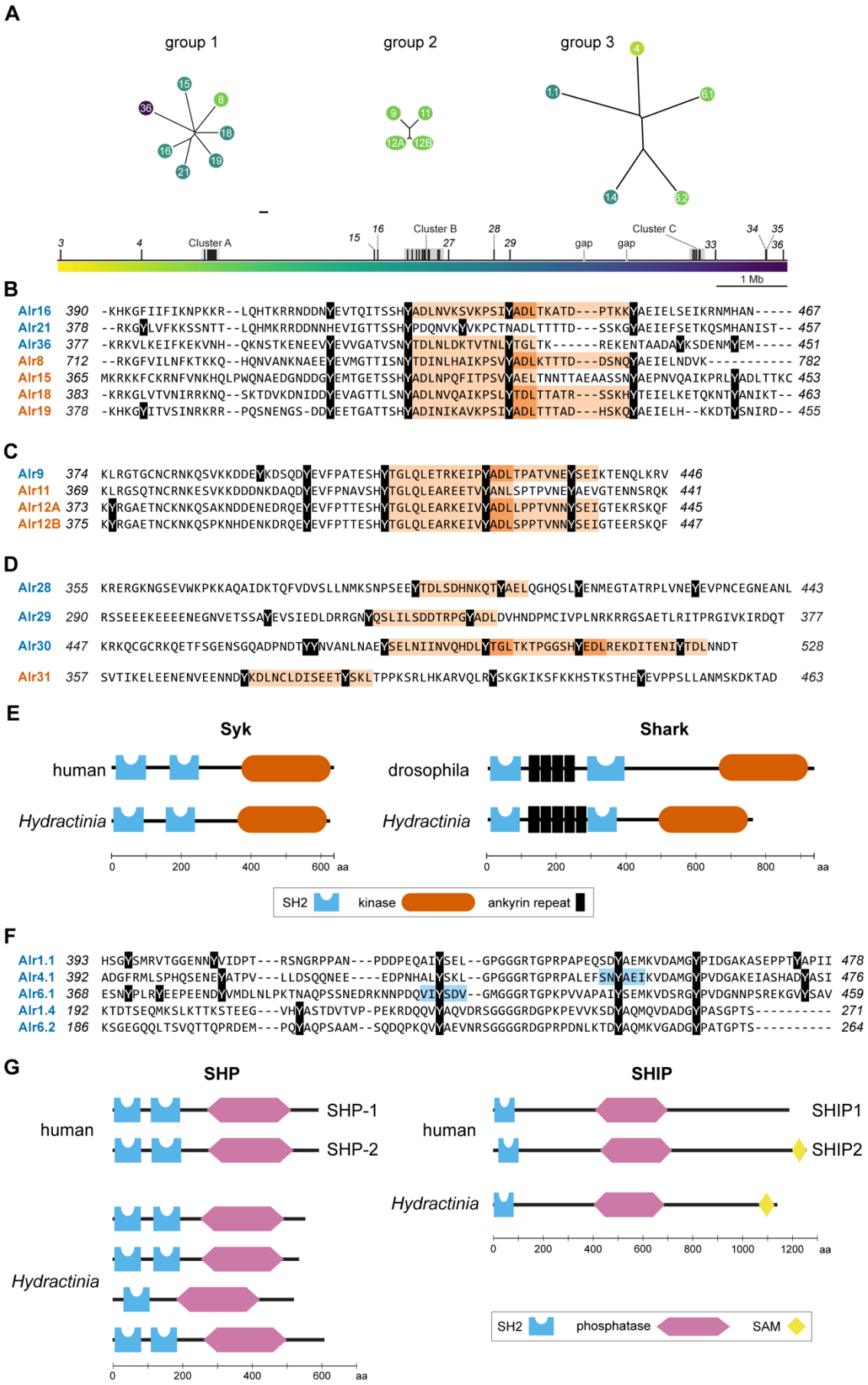

Fig. 4. Alr cytoplasmic tails sequence diversity and motifs.

(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) ITAMs in ungrouped cytoplasmic tails. In (B-D), *bona fide* gene names are in blue, putative gene names are in orange. ITAMs have an orange background. Darker orange indicates overlapping ITAMs.

(E) Comparison of human Syk and *Drosophila* Shark to *Hydractinia* Syk and Shark.

(F) Alignment of group 3 cytoplasmic tails. ITIMs have a blue background. All tyrosines have a black background. In (B-E) the alignment is truncated to highlight the ITIM/ITAM motifs, and all tyrosines are shown with a black background.

(G) Comparison of human SHP and SHIP proteins to *Hydractinia* SHP and SHIP homologs.

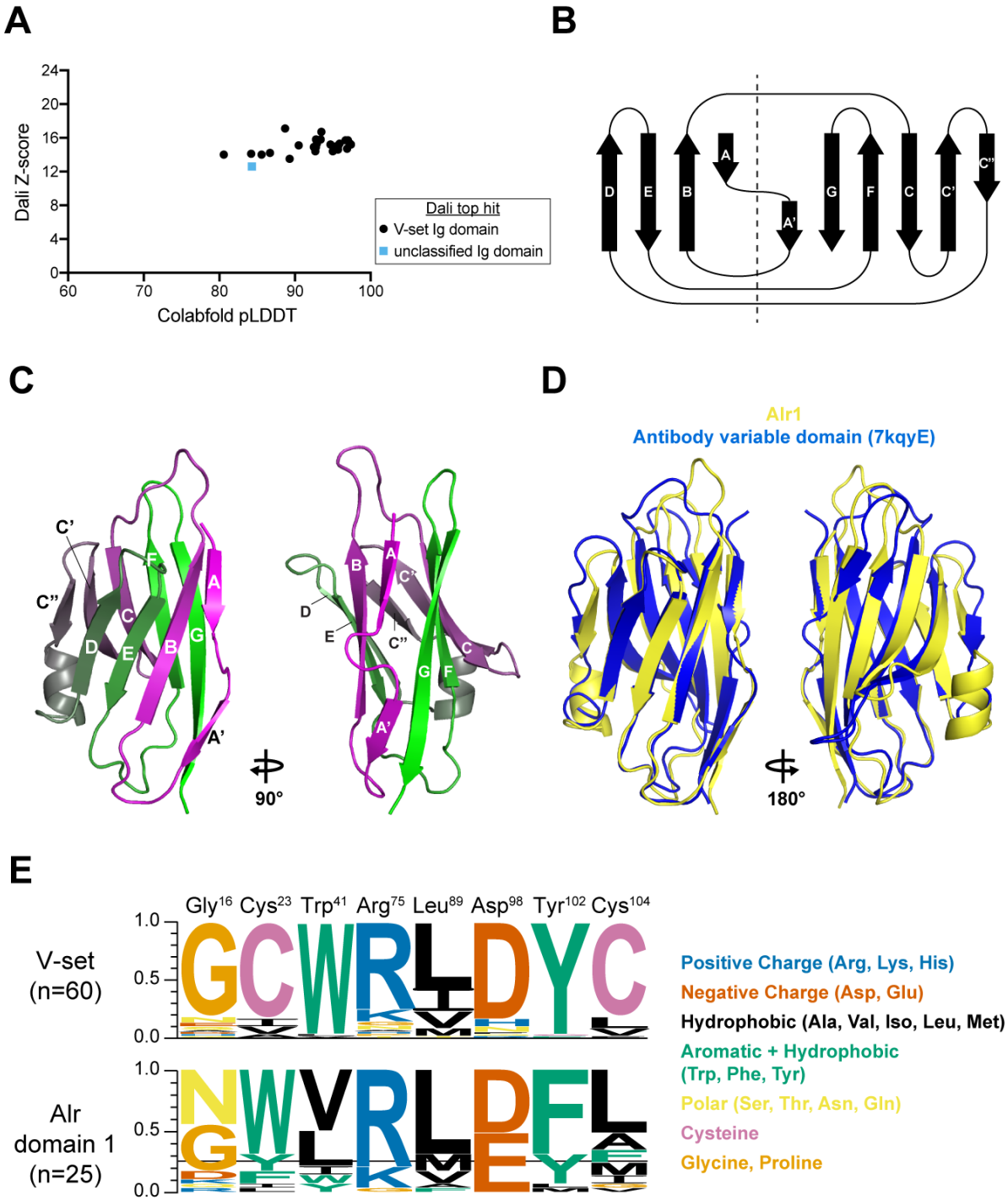

Fig. 5. Sequence analysis and structural predictions of domain 1.

(A) Plot of each domain 1 model's average pLDDT score versus its alignment Z-score to the top PDB model identified by DALI.

(B) Structural alignment of Alr1 domain 1 to a V-set Ig domain from human heavy-chain antibody (PDB 7KQY).

(C) Topology of β-strands in V-set folds. Dotted line separates the strands within the same β-sheet.

(D) V-set β-strands labeled on the predicted structure of Alr1 domain 1.

(E) Sequence logo comparing frequency of amino acids at eight conserved positions in V-set Ig domains (top) and Alr domain 1 sequences (bottom).

Figure 6 (Legend on next page)

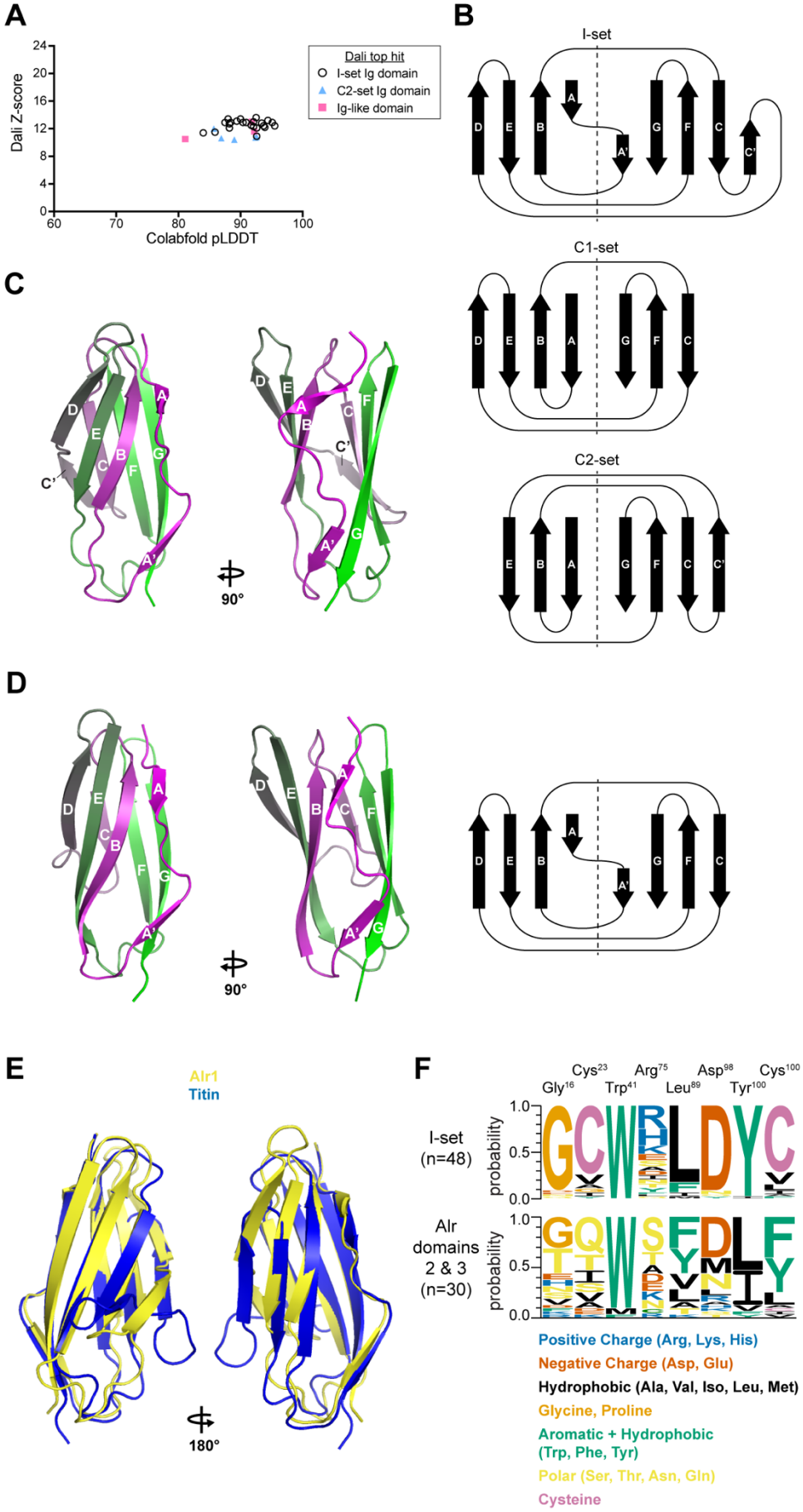

Fig. 6. Sequence analysis and structural predictions of domains 2 and 3.

(A) Plot of each domain 2 and domain 3 model's average pLDDT score versus its alignment Z-score to the top PDB model identified by DALI.

(B) Topology of β-strands in I-set, C1-set, and C2-set folds. Strand D is part of the EBA sheet in C1 Ig-domains, but is part of the CFG sheet in Fn3 and C2-set domains, where it is often labeled C'. Dotted line separates the strands within the same β-sheet.

(C) I-set β-strands labeled on the predicted structure of Alr1 domain 2.

(D) β-strands labeled on the predicted structure of Alr2 domain 2. The corresponding Greek key is shown to the right.

(E) Structural alignment of Alr1 domain 2 to an I-set Ig domain from titin (PDB 2RJM).

(F) Sequence logo comparing frequency of amino acids at conserved positions in I-set Ig domains (top) and Alr domains 2 and 3 (bottom).

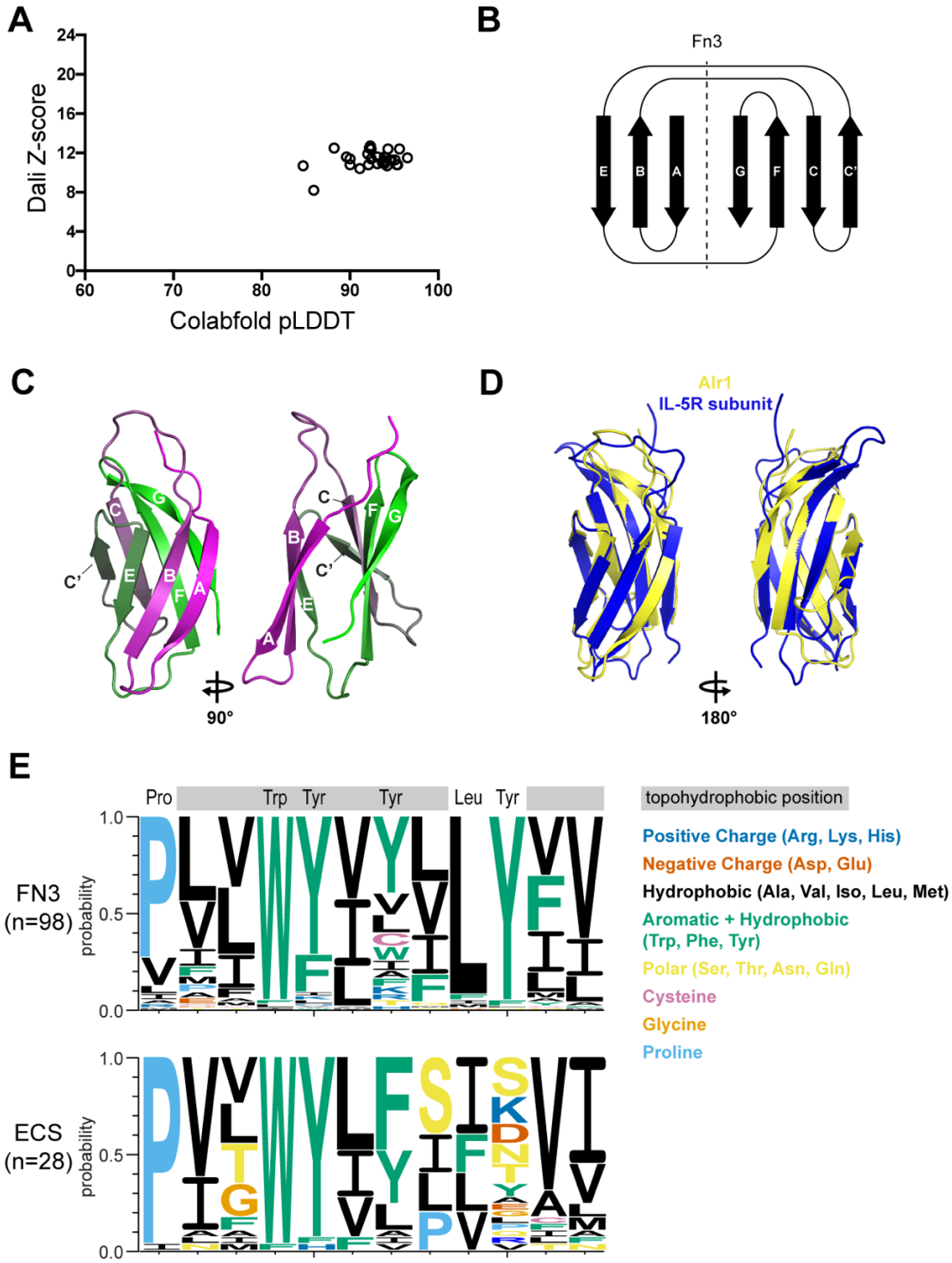

Fig. 7. Sequence analysis and structural predictions of the ECS.

(A) Plot of each ECS model's average pLDDT score and corresponding alignment Z-score to the top PDB model identified by DALI.

(B) Topology of β-strands in Fn3 folds. Dotted line separates the strands within the same β-sheet,

(C) Predicted topology of β-strands in the Alr1 ECS folds.

(D) Structural alignment of Alr1 ECS to a Fn3 domain from Interleukin-5 receptor subunit alpha (PDB 6H41).

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

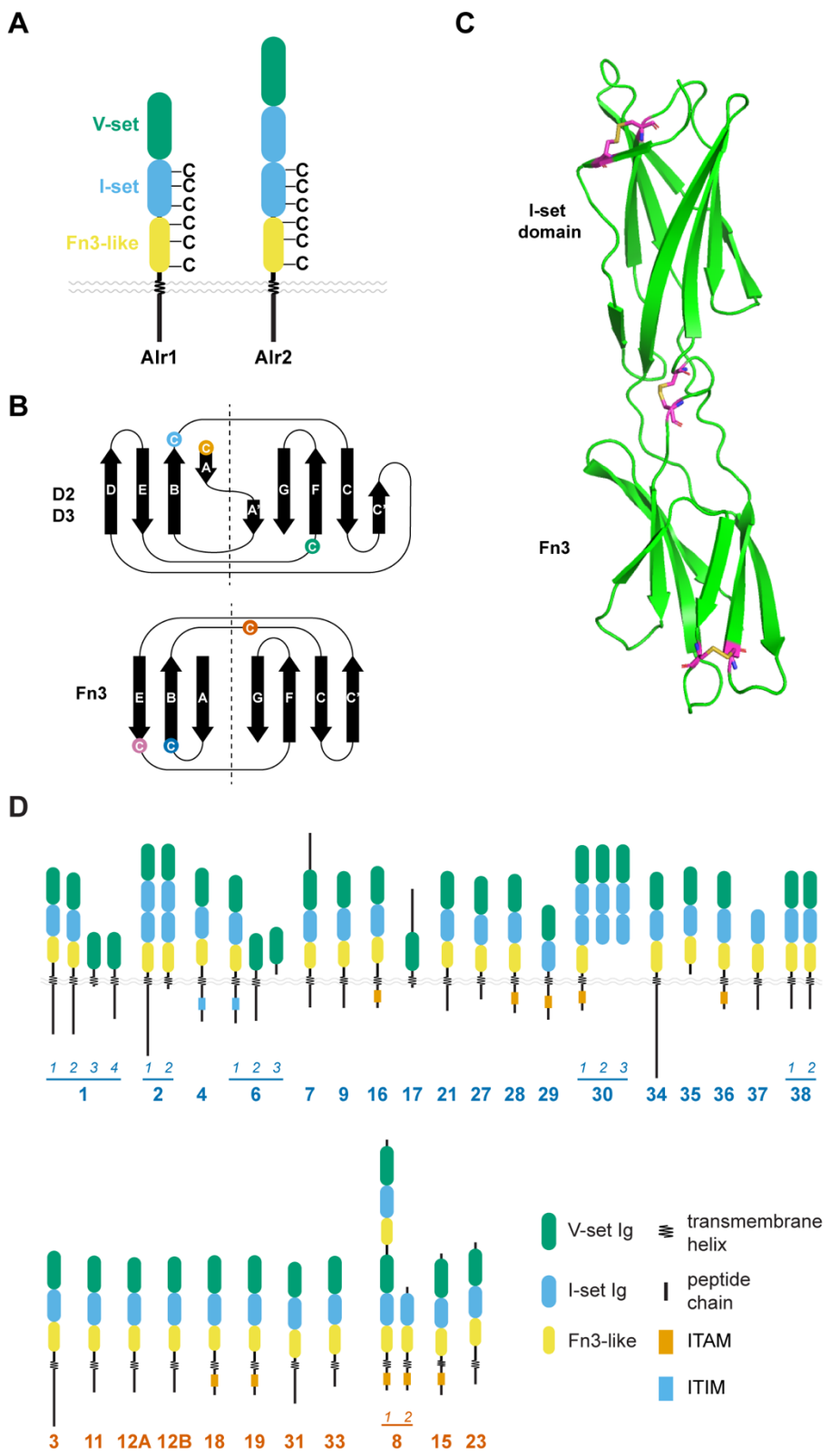

Fig. 8. Invariant cysteines in Alr proteins and final domain predictions.
(A) Occurrence of invariant cysteine residues found in Alr proteins with an Alr1-like or Alr2-like domain architecture.
(B) Position of invariant cysteines in domains 2, 3 (I-set), and the ECS fold (Fn3-like).
(C) Structural prediction of tandem I-set and Fn3-like domains of Alr1 shows each invariant cysteine forms a disulfide bond.
(D) Domain architecture and motifs of *bona fide* and putative Alr proteins

LIST of SUPPLEMENTARY MATERIALS

Fig. S1. Pedigree of colonies used to generate ARC-F reference sequence.
Fig. S2. Detail of ARC reference assembly.
Fig. S3. Alternative splicing at Alr30 and Alr2
Fig. S4. Alignment of Alr cytoplasmic tails
Fig. S5. STRIDE secondary structure predictions for Alr domain 1.
Fig. S6. Multiple sequence alignment of V-set Ig domains and Alr domain 1.
Fig. S7. Measurements of the turn from strands A' to B in the predicted structures of domain 1.
Fig. S8. STRIDE secondary structure predictions for Alr domain 2 and domain 3.
Fig. S9. Multiple sequence alignment of I-set Ig domains and Alr domains 2 and 3.
Fig. S10. Measurements of the turn from strands A' to B in the predicted structures of domains 2 and 3.
Fig. S11. Sequence logo of I-set-specific motifs in classical I-set domains and Alr domains 2 and 3.
Fig. S12. Sequence removed from the ECS prior to structural prediction.
Fig. S13. STRIDE secondary structure predictions for Alr ECS fold.
Fig. S14. Multiple sequence alignment of Fn3 domains and Alr ECS fold.
Table S1. Overlap coordinates of genomic contigs and BAC contigs used to create reference ARC-F sequence.
Table S2. Gene models classified as Alr pseudogenes
Table S3. Sequence homology of Domain 1
Table S4. Predicted structural homology for domain 1
Table S5. Sequence homology for Domain 2 and 3
Table S6. Predicted structural homology for domains 2 and 3
Table S7. Sequence homology of the ECS fold
Table S8. Predicted structural homology for the immunoglobulin-like fold in the ECS
Table S9. Structural predictions of tandem I-set and FnIII-like domains
Table S10. BLASTP results for full-length Alr protein sequences
File S1. FASTA formatted sequence of the ARC-F reference. Two gaps of unknown physical size are denoted with N's.
File S2. GFF3-formatted annotations of *Alr* genes in the ARC-F reference sequence.
File S3. FASTA formatted sequence of contig utg7180000000456, which contains *Alr37*.
File S4. GFF3-formatted annotation of *Alr37* on contig utg7180000000456.
File S5. FASTA-formatted sequence of contig utg7180000000115, which contains *Alr38*.
File S6. GFF3-formatted annotation of *Alr38* on contig utg7180000000115
File S7. FASTA-formatted cDNA sequences of bona fide genes in the *Alr* gene family.
File S8. FASTA-formatted amino acid sequences of Alr proteins encoded by bona fide genes.
File S9. FASTA-formatted cDNA sequences of putative genes in the *Alr* gene family.
File S10. FASTA-formatted amino acid sequences of Alr proteins encoded by putative genes.
File S11. FASTA-formatted MAFFT alignment of amino acid sequences for domain 1 from *bona fide* and putative *Alr* genes.
File S12. FASTA-formatted MAFFT alignment of amino acid sequences for domains 2 and 3 from *bona fide* and putative *Alr* genes.

File S13. FASTA-formatted MAFFT alignment of amino acid sequences for the ECS from *bona fide* and putative *Alr* genes.

File S14. Zip file containing PDB files of the top predicted Colabfold models for all *bona fide* and putative domain 1, domains 2 and 3, and the trimmed ECS.

File S15. FASTA-formatted amino acid sequences of the trimmed ECS used for structural predictions and alignment to fibronectin III domains.