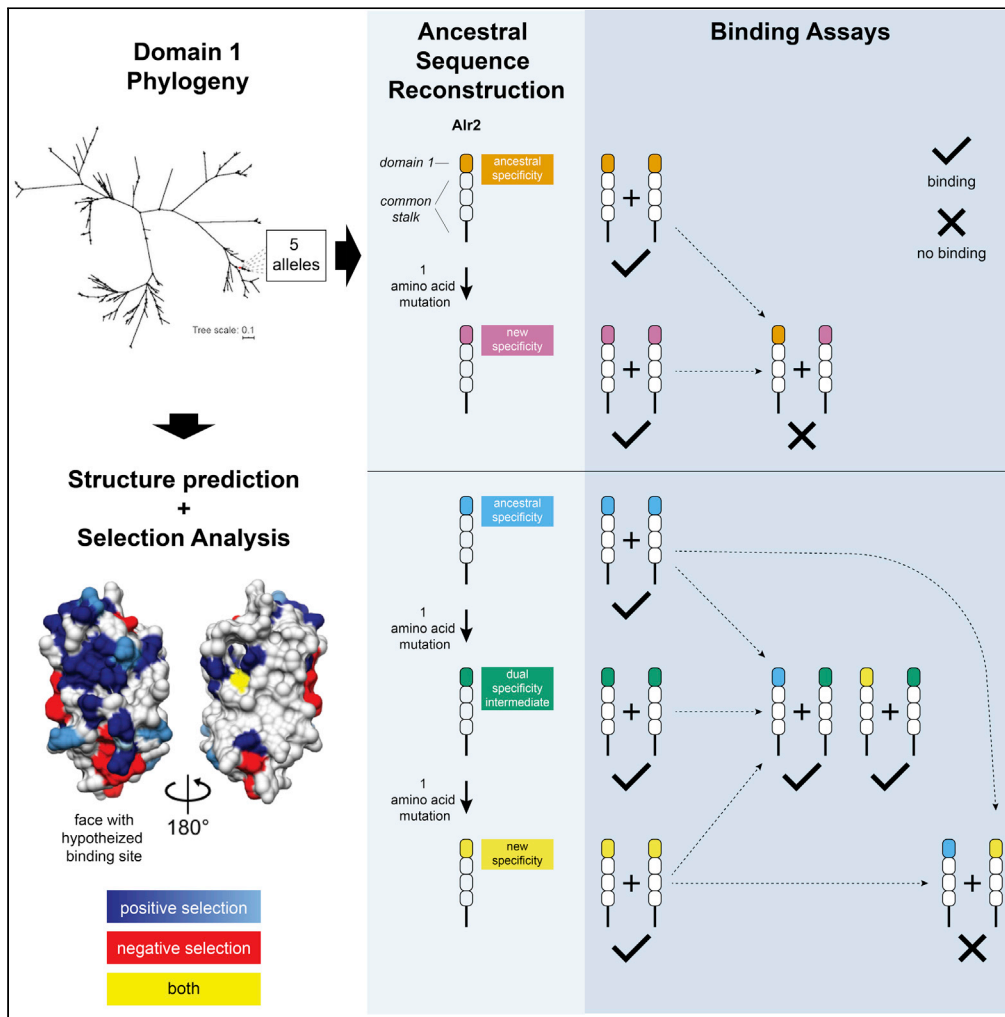## Article

New binding specificities evolve via point mutation in an invertebrate allorecognition gene

Aidan L. Huene,
Traci Chen,
Matthew L.
Nicotra

matthew.nicotra@pitt.edu

Highlights

Three binding specificities evolved in a clade of five domain 1 sequences

One new specificity evolved via a single amino acid mutation

Another new specificity evolved through a dual-specificity intermediate

Sequence analyses suggest a possible binding interface

Article

New binding specificities evolve via point mutation in an invertebrate allorecognition gene

Aidan L. Huene,[1,2] Traci Chen,[1] and Matthew L. Nicotra[1,2,3,4,5,*]

SUMMARY

Many organisms use genetic self-recognition systems to distinguish themselves from conspecifics. In the cnidarian, *Hydractinia symbiolongicarpus*, self-recognition is partially controlled by *allorecognition 2* (*Alr2*). *Alr2* encodes a highly polymorphic transmembrane protein that discriminates self from nonself by binding *in trans* to other Alr2 proteins with identical or similar sequences. Here, we focused on the N-terminal domain of Alr2, which can determine its binding specificity. We pair ancestral sequence reconstruction and experimental assays to show that amino acid substitutions can create sequences with novel binding specificities either directly (via one mutation) or via sequential mutations and intermediates with relaxed specificities. We also show that one side of the domain has experienced positive selection and likely forms the binding interface. Our results provide direct evidence that point mutations can generate Alr2 proteins with novel binding specificities. This provides a plausible mechanism for the generation and maintenance of functional variation in nature.

INTRODUCTION

The ability to discriminate self from same-species nonself (often referred to as allorecognition) has evolved in plants (Fujii et al., 2016), fungi (Paoletti, 2016; Gonçalves et al., 2020), slime molds (Kundert and Shaulsky, 2019), marine invertebrates (Nicotra, 2019), and bacteria (Gibbs and Greenberg, 2011; Pathak et al., 2013; Cao et al., 2019). In all cases, it is based on an organism's genotype at polymorphic loci. This polymorphism is thought to be maintained by a form of balancing selection called negative frequency-dependent selection (Wright, 1939; Kimura and Crow, 1964). Under negative frequency-dependent selection, alleles become more fit as they become less frequent. This is because rare alleles are unlikely to be shared by chance, making them better markers of self. New alleles, the rarest of all, spread in a population until their frequencies reach that of other alleles (Richman and Kohn, 2000). These dynamics can maintain tens to hundreds of self-recognition alleles in a population (Casselton and Olesnicky, 1998; Lawrence, 2000; Gloria-Soria et al., 2012; James, 2015; Nydam et al., 2017; Goncalves et al., 2019). How new, functional self-recognition alleles are generated and ultimately contribute to this extreme polymorphism remains a puzzle.

Hydractinia symbiolongicarpus is a colonial cnidarian that uses proteins that are their own ligand for allorecognition (Frank et al., 2020). *Hydractinia* colonies begin when a sexually produced larva settles on a hermit crab shell and metamorphoses into a polyp. The animal then expands across the shell by elongating stolons (extensions of its gastrovascular system) or mat (a plate of tissue that fills the space between stolons), from which new polyps grow to form a mature colony. As it grows, a colony's stolons and mat edges meet and fuse to create an anastomosing network of gastrovascular canals embedded in a continuous sheet of mat. The colony will also fuse to itself as it grows around the shell or recovers from injury. Because nearly half of all shells bear more than one colony (Yund et al., 1987), colonies also frequently encounter conspecifics. This usually elicits an aggressive rejection response in which the colonies fight by firing nematocysts (harpoon-like organelles) until one dies (Nicotra and Buss, 2005).

Previous experiments with inbred, laboratory strains of *Hydractinia* have demonstrated that colonies can distinguish self from nonself by their genotype at two linked genes called *allorecognition 1 (Alr1)* and *allorecognition 2 (Alr2)* (Cadavid et al., 2004; Powell et al., 2007, 2011). Animals that shared at least one allele at both loci fused, while those that shared no alleles at either *Alr1* or *Alr2* rejected. If colonies only shared alleles at one locus, they fused but then separated. Because only two alleles were present at each locus in these strains, it was impossible to determine how similar alleles need to be for colonies to

[1]Thomas E. Starzl Transplantation Institute, Department of Surgery, University of Pittsburgh, Pittsburgh, PA 15260, USA

[2]Center for Evolutionary Biology and Medicine, University of Pittsburgh, Pittsburgh, PA 15260, USA

[3]Department of Immunology, University of Pittsburgh, Pittsburgh, PA 15260, USA

[4]Twitter: @mattnicotra

[5]Lead contact

*Correspondence:
matthew.nicotra@pitt.edu

https://doi.org/10.1016/j.isci.2021.102811

fuse. In addition, subsequent experiments with wild-type colonies have strongly suggested at least one additional allorecognition locus exists in the genomic region encoding *Alr1* and *Alr2* (Powell et al., 2007, 2011; Nicotra et al., 2009; Rosa et al., 2010).

Alr1 and *Alr2* both encode type I transmembrane proteins with tandem Ig-like domains in their extracellular regions (Nicotra et al., 2009; Rosa et al., 2010). Each Alr is capable of cell-to-cell (i.e., *trans*) homophilic binding (Karadge et al., 2015). Binding is restricted to isoforms with identical or very similar sequences (Karadge et al., 2015). These results, combined with the fact that *Hydractinia* must share *Alr1* and *Alr2* alleles to recognize each other as self, have led to the hypothesis that homophilic binding of Alr1 and Alr2 between colonies is part of the *in vivo* self-recognition mechanism.

Alr1 and *Alr2* are also highly polymorphic. A study of *Alr2* identified 183 distinct Alr2 amino acid sequences from a single population (Gloria-Soria et al., 2012). *Alr1* is expected to be similarly diverse based on the extreme levels of sequence polymorphism observed in 20 sequenced alleles (Rosa et al., 2010). These observations suggest hundreds of distinct binding specificities could exist in nature.

Two features of *Hydractinia*'s natural history likely contribute to the evolution of this extreme polymorphism. First, colonies must be able to compete for space while simultaneously retaining the ability to recognize and fuse to themselves. Thus, a new allele that binds only to itself is favored because it permits a colony to compete with every other *Hydractinia* in the population but still fuse with itself. Second, *Hydractinia* has a pluripotent stem cell lineage that can differentiate into germ cells at any point in the colony's life. Fusion allows these stem cells to migrate from one colony into the other, where they could dominate its gametic output. This phenomenon, called stem cell parasitism, has been observed anecdotally in *Hydractinia* (Künzel et al., 2010; Dubuc et al., 2020) and is thought to be a common trait in most colonial organisms (Buss, 1987; Stoner and Weissman, 1996; Stoner et al., 1999; Laird et al., 2005; Aanen et al., 2008). Thus, a new allele that restricts fusion to self would be favored because it would reduce the risk of stem cell parasitism.

It has been assumed that novel *Alr1* and *Alr2* alleles are generated by random mutations that are then subjected to negative frequency-dependent selection. This raises the question of whether point mutations, by themselves, can generate alleles with novel homophilic binding specificities and, furthermore, whether this type of mutation could, in part, explain the large number of binding specificities thought to exist in natural populations.

Here, we sought to determine how binding specificities evolve in the N-terminal domain of Alr2. This domain, referred to as "domain 1," is the most polymorphic region of Alr2. Changes in domain 1 can prevent Alr2 proteins from binding and therefore might be able to generate alleles with new identities. To determine how this domain has evolved in nature, we identified a clade of five domain 1 sequences encoding isoforms that differed by six or fewer amino acids. We then used ancestral sequence reconstruction and *in vitro* binding assays to determine the evolutionary history of the clade. Our results demonstrate that the binding specificity of domain 1 can be altered by single amino acid changes, resulting in novel specificities or intermediates with broadened specificities. Finally, we show that one face of the predicted domain 1 structure appears to be under diversifying selection, which also allows us to hypothesize that Alr2 protein-protein interactions occur in a side-to-side manner.

RESULTS

Point mutations in domain 1 can create new binding specificities

We searched a data set of full-length, naturally occurring *Alr2* alleles (Nicotra et al., 2009; Gloria-Soria et al., 2012) and identified two (111A06 and 214 × 10^6) that encoded Alr2 allelic isoforms (hereafter, "isoforms") with six amino acid differences in domain 1 and identical sequences across the rest of the extracellular region (Figures 1A and 1B). Using cell aggregation assays (Karadge et al., 2015), we found that each isoform bound to itself across opposing cell membranes but did not bind to the other (Figure 1C). We therefore sought to identify the amino acid differences that prevented them from binding to each other.

Each amino acid difference between 111A06 and 214E06 is the result of one point mutation. To reconstruct the evolutionary history of these mutations, we created a phylogeny of all known domain 1 coding sequences (Figure 2A). 111A06 and 214E06 were located in a clade with three additional sequences
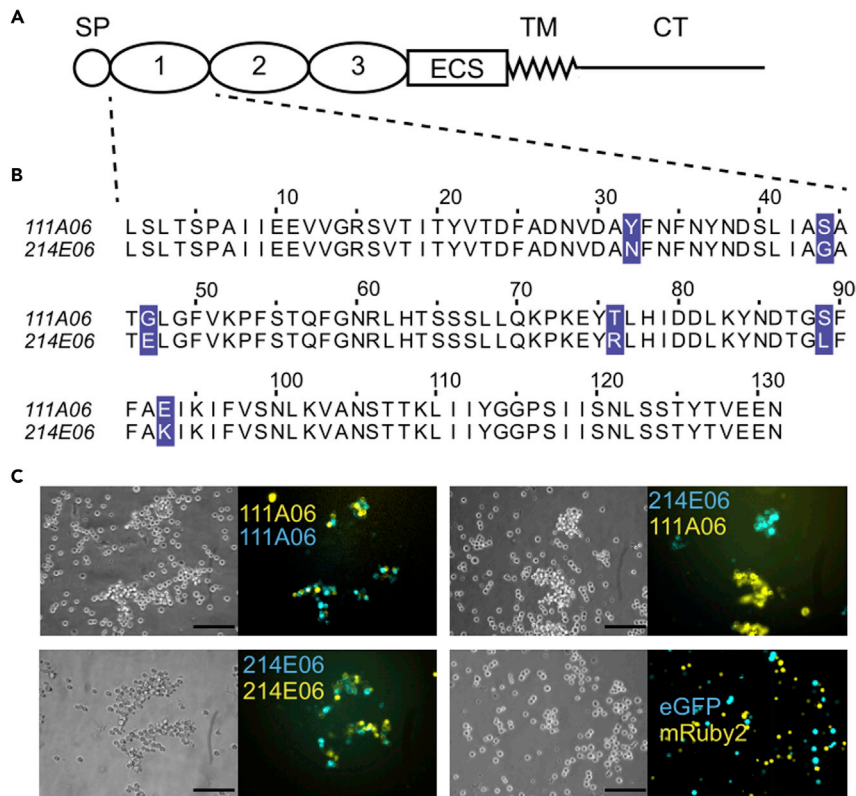
Figure 1. Isoform-specific, homophilic binding of Alr2 isoforms
(A) Alr2 protein structure. SP = Signal peptide, ECS = Extracellular spacer, TM = Transmembrane domain, CT = Cytoplasmic tail.
(B) Multiple sequence alignment of 111A06 and 214E06 domain 1. Polymorphisms highlighted in purple.
(C) Cell aggregation assays of 111A06 and 214E06. Cells transfected with vectors encoding only fluorescent proteins (eGFP or mRuby2) do not form aggregates (bottom right). Scale bar = 100 μm.

(Figure 2B). We then used ancestral sequence reconstruction to infer the sequence of each node. All but the ancestral node (Anc) were predicted to be identical to an extant sequence (Figure 2C). Because 214E06 and Hap010 differed only by a single synonymous mutation, we used 214E06 to represent their shared amino acid sequence.

To determine the binding specificity of the domain 1 isoforms encoded by these sequences, we expressed each as a fusion to domain 2 through the cytoplasmic tail of the 111A06 isoform, with a C-terminal fluorescent protein tag (Figure 2D). The resulting isoforms were tested against themselves and each other in cell aggregation assays (Figures 2E–2H). Each isoform, including the predicted ancestor, Anc, caused cells to form multicellular aggregates, indicating it was capable of homophilic binding (Figure 2E). In pairwise assays, 111A06 did not form mixed aggregates with any isoform, indicating it had a unique binding specificity within the clade (Figure 2F). In contrast, Anc, 046B, and Hap074 all formed mixed aggregates with each other, indicating a shared binding specificity (Figure 2G).

In assays that paired 214E06 with Anc, 046B, or Hap074, we observed single-color aggregates, some of which appeared to adhere to aggregates of a different color (Figure 2H, arrowheads). These semimixed aggregates were repeatable (Figure S1A) and qualitatively different from the mixed aggregates it formed when paired with itself and from the completely separate aggregates it formed with the other four isoforms. This ruled out a defect in 214E06 that prevented homophilic binding or caused it to bind to any isoform. Semimixed aggregates have been observed in studies of cell adhesion molecules that have strong homophilic affinities but weaker heterophilic affinities (Katsamba et al., 2009; Goodman et al., 2016). Because of this, we concluded that 214E06 binds more weakly to Anc, 046B, and Hap074 than to itself and that it therefore had a different binding profile from the other isoforms.
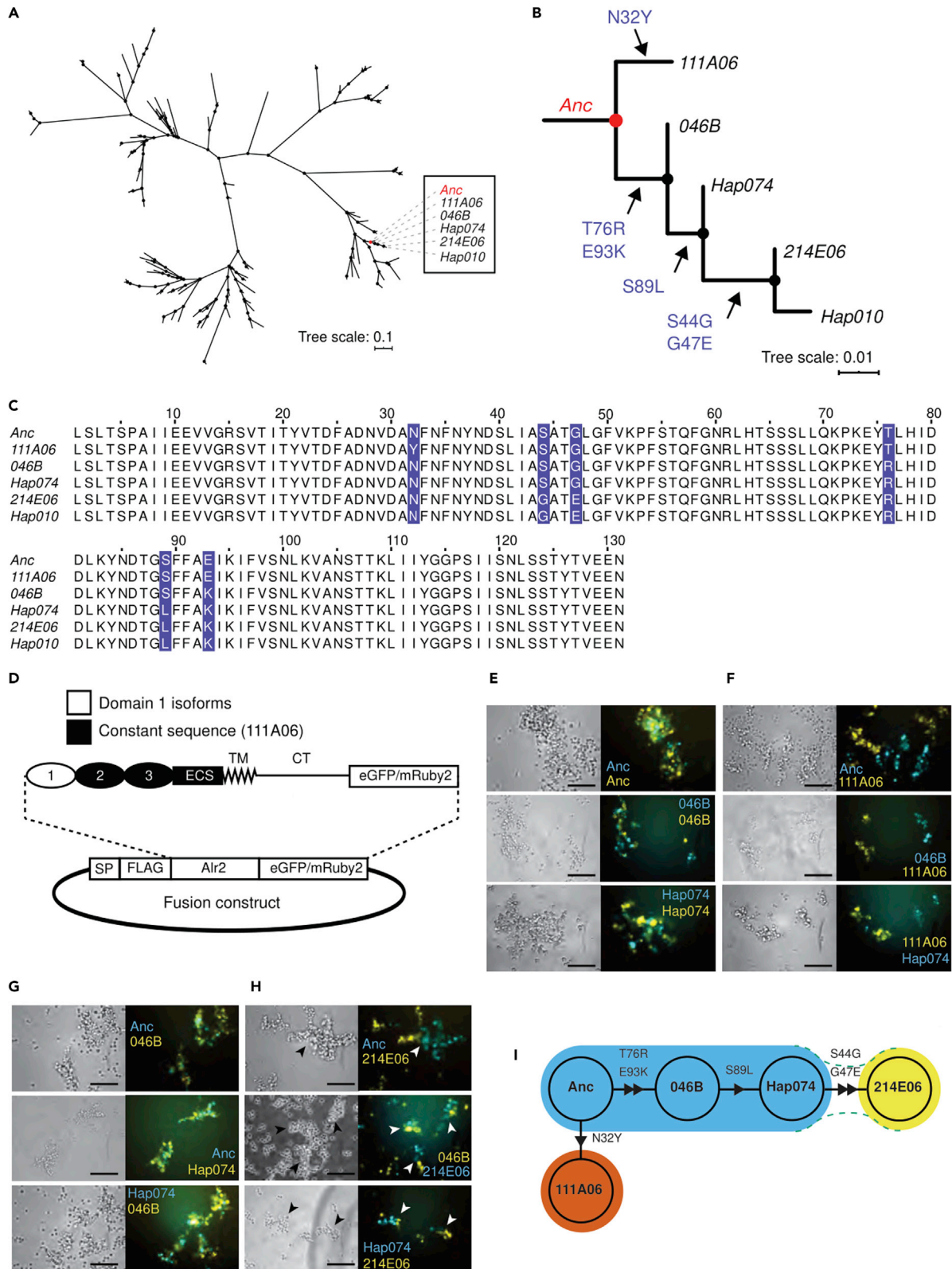
Figure 2. Evolution of novel binding specificities via point mutation

(A) Maximum-likelihood tree of 146 domain 1 coding sequences.

(B) Expansion of clade that includes 111A06 and 214E06. Allele names on branch. Amino acid changes indicated along branches.

Figure 2. *Continued*

(C) Multiple sequence alignment of clade. Variant residues highlighted.

(D) Plasmid map Alr2 fusion proteins. (E-H) Representative images of cell aggregation assays.

(E) Anc, 046B, and Hap074 against themselves.

(F) Anc, 046B, and Hap074 against 111A06.

(G) All pairwise combinations of Anc, 046B, and Hap074.

(H) Anc, 046B, and Hap074 versus 214E06. Arrowheads point to semi-mixed aggregates (See also Figure S1). Scale bar = 100 μm.

(I) Node network of isoforms colored by binding specificity. Triangles indicate the hypothesized direction of mutation from Anc. Green dotted lines indicate weaker heterophilic interactions.

Our results are consistent with the following evolutionary history (Figure 2I). An ancestral sequence, Anc, underwent a single mutation, N32Y, which created a daughter sequence, 111A06, with a novel binding specificity. In a separate lineage, the Anc sequence underwent two mutations, T76R and E93K to create 046B, which retained the ability to bind to Anc. A third mutation, S89L, then created Hap074, which also remained able to bind Anc and 046B. Two more mutations, S44G and G47E, then created 214E06, which bound more weakly to the ancestral isoforms than to itself (Figure 2I, dotted lines). The result is a clade in which we can discern three binding specificities, one of which arose via a single-point mutation.

New homophilic specificities can evolve via less restricted intermediates

Within the phylogeny, two pairs of mutations occurred within single branches (Figure 2B), preventing us from determining which came first. To determine whether the missing single-step intermediates were functional (i.e., able to bind homophilically) or had a different binding specificity from their parent and daughter sequences, we recreated each one (Figure 3A) and tested it in cell aggregation assays. We found each intermediate could bind homophilically (Figure 3B), thus ruling out the possibility that there were nonfunctional intermediates in the clade.

We next tested the specificity of each missing intermediate. The first pair, Anc-T76R and Anc-E93K, formed mixed aggregates with Anc, 046B, and Hap074 (Figure 3C). Assays pairing Anc-T76R with 111A06 resulted in single-color aggregates (Figure 3D), but those pairing Anc-E93K with 111A06 resulted in a few semi-mixed aggregates (Figure 3D, arrowheads, Figure S1B). Both mutants also formed semimixed aggregates when paired with 214E06 (Figures 3E and S1C). Thus, evolution from Anc to 046B is unlikely to have involved a significant change in binding specificity (Figure 3A).

In contrast, the specificity of the second pair of intermediates, Hap074-S44G and Hap074-G47E, was different from their parent and daughter sequences. These mutants failed to form mixed aggregates with 111A06 (Figure 3F) but did form mixed aggregates with 214E06 (Figure 3G) and all other ancestral sequences (Figure 3H). We did not observe semimixed aggregates in any assay. These results suggest the first mutation on the path from Hap074 to 214E06, either S44G or G47E, created a sequence that could still bind Hap074 (Figure 3A). The acquisition of the second mutation then generated a new allele, 214E06, which remained able to bind its parent sequence, but had a weaker affinity for Hap074. The evolution of new domain 1 sequences can therefore proceed through intermediates with broader specificities than their parental or daughter sequences.

The N32Y mutation preserves homophilic binding and alters specificity

Isoform 111A06 evolved when position 32 mutated from Asn to Tyr in Anc. We therefore hypothesized the N32Y mutation might turn 046B or Hap074, which had the same specificity as Anc, into isoforms with the same specificity as 111A06. To test this, we generated 046B-N32Y and Hap074-N32Y. In assays with themselves, each formed mixed aggregates, indicating the mutation did not disrupt homophilic binding (Figure S1D). In pairwise assays with each other and 111A06, the mutants formed mixed aggregates, indicating they had gained the ability to bind 111A06 and each other (Figures 4A and S1G). In pairwise assays with their immediate ancestors, however, the mutants formed semimixed aggregates (Figure 4B, asterisks, and Figure S1E). This indicated each could still bind its ancestor, albeit more weakly than it did itself. Finally, we performed pairwise assays with the remaining isoforms in the clade. This showed the mutants had different specificities than 111A06, 046B, or Hap074 (Figures 4B and S2). In sum, the N32Y altered the specificities of 046B and Hap074 but did not generate daughter sequences with the same specificity as 111A06.

We next tested whether the N32Y mutation would alter the specificity of 214E06, the remaining domain 1 isoform known to exist in nature. We generated 214E06-N32Y and found it formed mixed aggregates with
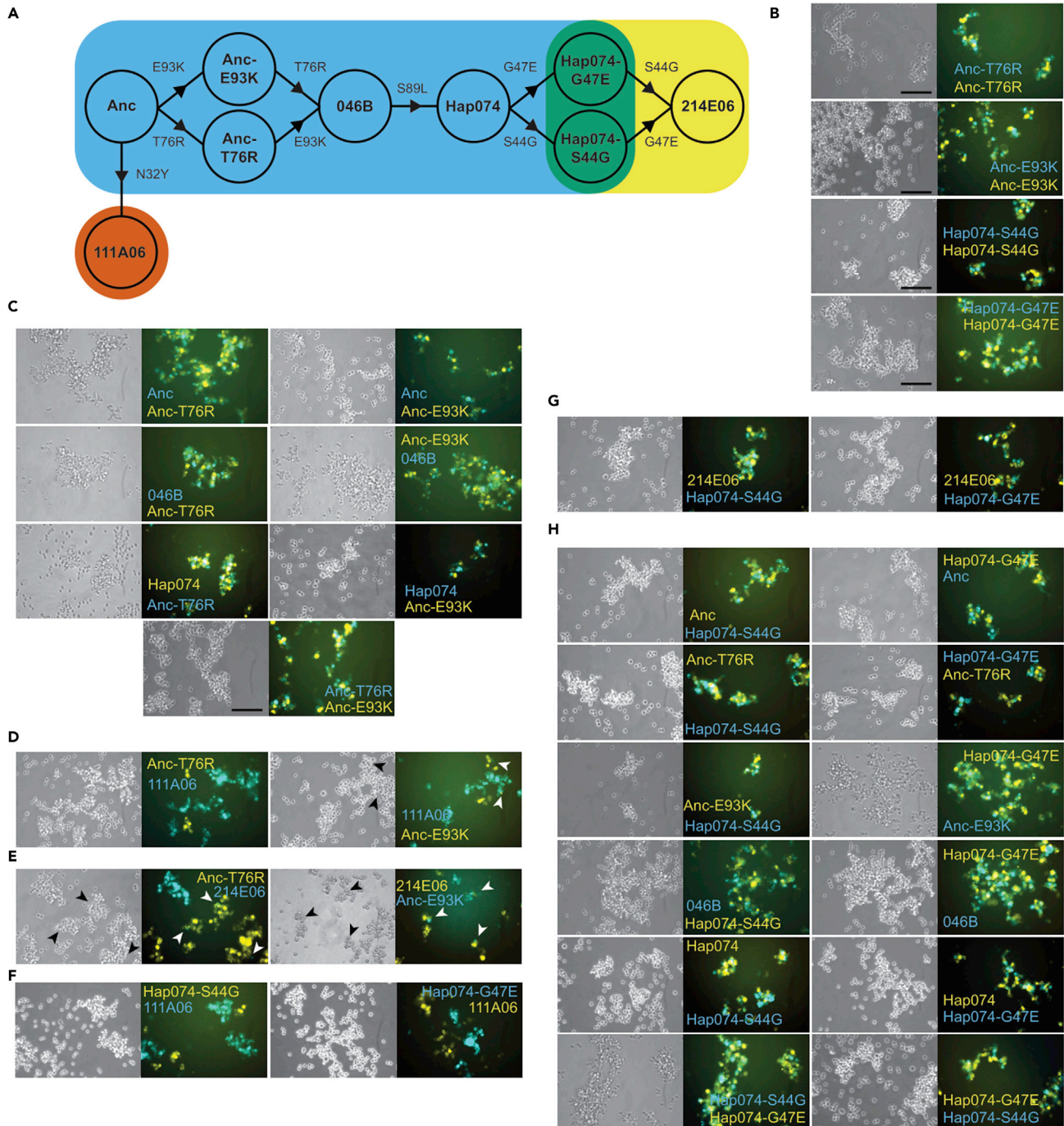
Figure 3. Domain 1 isoforms can evolve via intermediates with broadened specificity

(A) Expanded node network including hypothesized single-step mutants between Anc and 046B, Hap074 and 214E06. Scale bar = 100 μm and applies to all images.

(B–H) Representative images of cell aggregation assays.

(B) Mutants tested against themselves.

(C) Anc-T76R and Anc-E93K tested against Anc, 046B, Hap074.

(D) Anc-T76R versus 111A06 (left) and Anc-E93K versus 111A06 (right). Semi-mixed aggregates indicated with arrowheads (See also Figure S1A).

(E) Anc-T76R and Anc-E93K versus 214E06 (See also Figure S1C).

(F) Hap074-S44G and Hap074-G47E versus 111A06.

(G) Hap074-S44G and Hap074-G47E versus 214E06.

(H) Hap074-S44G and Hap074-G47E versus all remaining isoforms.

Figure 4. Effects of N32Y mutation on binding specificity and structural analysis

(A) Results of assays between N32Y mutants and 111A06 (See also Figures S1G and S1H).

(B) Binding profiles of 111A06, 046B-N32Y, Hap074-N32Y, 046B, and Hap074. Asterisk denotes the result of an allele and its N32Y mutant.

(C) Binding profiles of 214E06 and 214E06-N32Y.

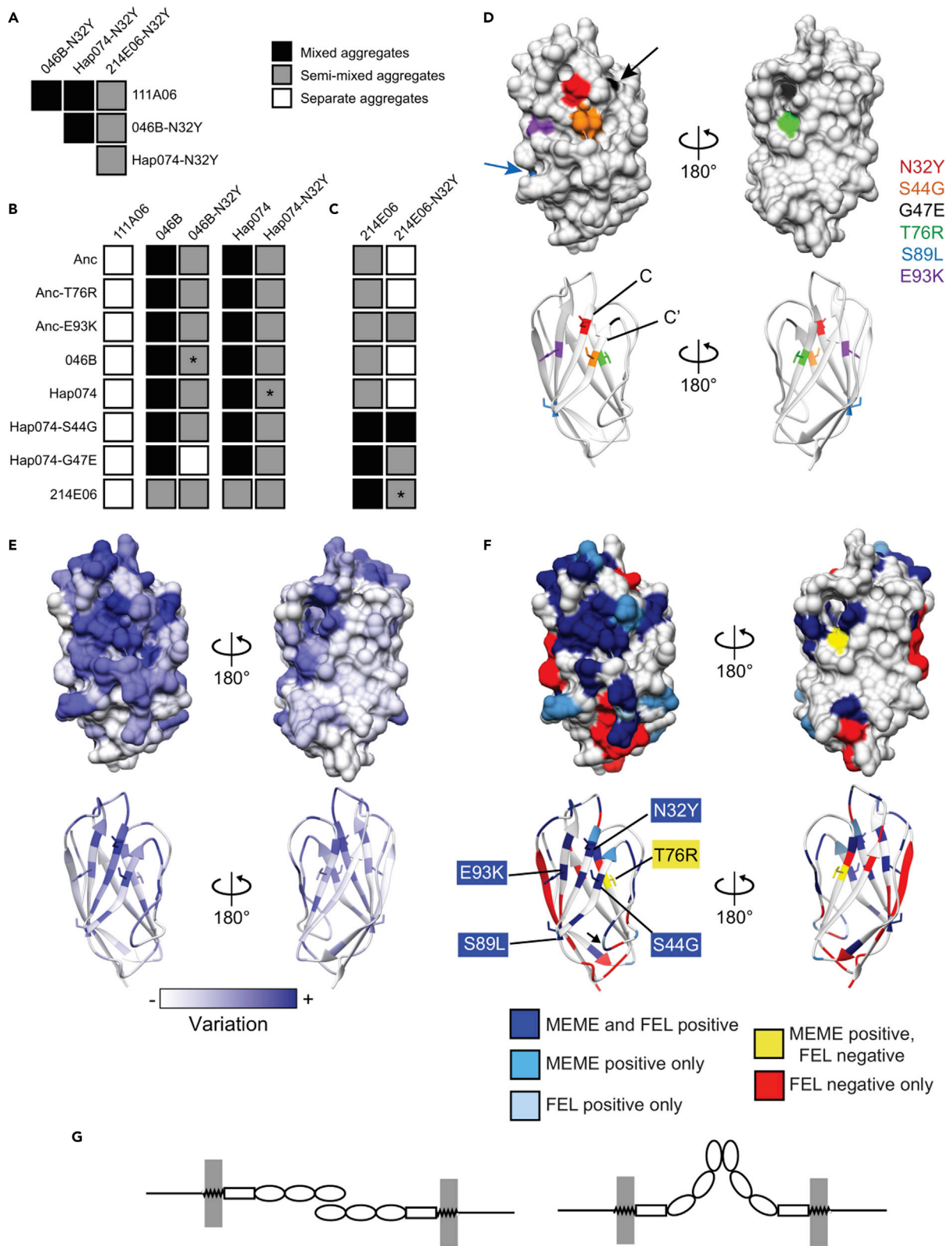(D) Predicted structure of Anc domain 1. Six variant residues labeled.

(E) Sequence conservation mapped onto domain 1.

(F) Residues predicted to have experienced either diversifying or purifying selection mapped onto domain 1. Colors correspond to the predictions of MEME and/or FEL. Arrowhead indicates the one residue predicted to be under positive selection by FEL only.

(G) Hypothetical binding topologies Alr2.

itself (Figure S1D), indicating it was able to bind homophilically to itself. It also formed semimixed aggregates with 214E06 (Figure 4C, asterisk, and Figure S1F), indicating a reduced binding affinity for its immediate ancestor compared to itself. However, 214E06-N32Y only formed semimixed aggregates with 111A06, 046B-N32Y, and Hap074-N32Y (Figure 4A). Thus, simply sharing a Tyr at position 32 was insufficient for isoforms to bind each other as strongly as they did themselves. Pairwise assays with the remaining isoforms revealed 214E06-N32Y to have a different binding profile than 214E06, with the exception of the mixed aggregates formed with Hap074-S44G (Figures 4C and S3). The effect of the N32Y mutation thus depends on the sequence context in which it occurs.

Structural and evolutionary analyses suggest a potential binding interface

In this study, three mutations changed the binding specificity of domain 1 (N32Y, S44G, and G47E), and three others did not. To investigate how these mutations might affect the tertiary structure of domain 1—and thus its binding specificity—we used I-TASSER to predict their structures. All were predicted to fold like V-set Ig-domains, which was consistent with previous work (Nicotra et al., 2009). Five mutations mapped to one face of the predicted beta-sandwich, with the three specificity-altering mutations in close proximity to each other in beta-strands C and C' (Figure 4D shows the structure of Anc for illustration). This suggested these strands are involved in homophilic binding between compatible domain 1 isoforms.

To gain further insight into the mechanism of homophilic binding, we compared the predicted structures of domains differing by a single amino acid (e.g., Anc vs 111A06). We noted many differences in the orientation of the mutated residues and their nearby amino acids. However, molecular dynamics simulations indicated these orientations were probably unstable (data not shown), so we did not analyze the models any further.

As an alternative approach to identify functionally important parts in domain 1, we reasoned that selection should increase sequence variation at or near the binding site. We therefore calculated the level of sequence variation at each site across all known domain 1 sequences, then mapped this metric onto the predicted structure of Anc. We found that most of the variable sites were also concentrated on the side of the domain that includes strands C and C' and residues 32, 44, and 47 (Figure 4E).

One explanation for this increase in variation is that positive (diversifying) selection is acting on amino acid positions at the binding interface because this can generate new specificities.

Although current sequence-based methods do not allow one to test whether a single mutation on a single-branch experienced positive selection (Murrell et al., 2012; Spielman et al., 2019), we were able to test whether positive selection has acted on specific sites in domain 1 across the entire phylogeny of domain 1 sequences. To do this, we analyzed the alignment of all known domain 1 sequences with MEME (Murrell et al., 2012) and FEL (Pond and Frost, 2005). Thirty sites were predicted to have experienced positive selection and were concentrated on the side of the domain that includes strands C and C' (Figure 4F, Table S1). Twenty sites were predicted to be under negative (purifying) selection and mapped to this side of the domain.

With respect to the six positions at which mutations occurred in our clade of interest, sites 32, 44, 89, and 93 were predicted to have experienced positive selection on at least one branch of the full phylogeny, but site 47 was not. Site 76 was predicted by MEME to be under positive selection, but by FEL to be under negative selection, a pattern consistent with a burst of diversifying selection against a background of purifying selection (Spielman et al., 2019). In all, these results are consistent with positive selection acting to increase sequence variation at sites on a probable binding face.

Taken together, these evolutionary signatures also suggest Alr2 proteins might bind via "side-to-side" interactions at their N-terminal domains. We speculate these interactions could occur in either an antiparallel or parallel topology (Figure 4G).

DISCUSSION

Domain 1 is the most polymorphic region of Alr2 (Gloria-Soria et al., 2012). Here, we demonstrate that sequence differences in this domain can prevent Alr2 isoforms from binding to each other. Then, by reconstructing the history of a small domain 1 sequence family, we show that new sequences capable of discriminating between themselves and their ancestors can evolve via point mutation. This can occur with as little as one mutation or via sequential mutations leading through intermediates with relaxed specificities. The fact that so few mutations occurred within this family also increases our confidence in our sequence reconstructions. Because sequence differences in domain 1 are sufficient to alter Alr2 specificity, these mutations may have generated *Alr2* alleles with novel identities. Moreover, because the sequences in this study were drawn from a single population, our results show that natural selection maintains ancestral sequences alongside one encoding new specificities. Thus, our results reveal a mechanism capable of generating, maintaining, and increasing the functional diversity of *Alr2*.

In this study, we failed to identify domain 1 sequences that could not bind homophilically. This is somewhat surprising because alleles incapable of homophilic binding might be expected to exist in nature. Colonies that are $Alr2^{a/null}$ (where *a* is an allele encoding a homophilic binding protein and *null* cannot bind homophilically) might be functionally equivalent to $Alr2^{a/a}$ colonies. This is possible because fusions between colonies sharing only one allele are identical to fusions between colonies that share two alleles. Colonies with null alleles might even have a fitness advantage because the probability that they will fuse with nonself is reduced from the sum of two allele frequencies to the frequency of a single allele. So, why have we not detected null alleles in this and a previous study (Karadge et al., 2015)? One possibility is that null alleles are rare, and we have not found one yet because we have only studied ~5% of sequence variation at *Alr2*. A second possibility is that $Alr2^{a/null}$ animals are not, in fact, equivalent to $Alr2^{a/a}$ animals. This might be true if *Alr2* has essential functions beyond self-recognition at the colony border. In fact, we suspect this is the case because *Alr2* is constitutively expressed from embryonic development through adulthood and across all tissues in a colony (Nicotra et al., 2009). *Alr2* might therefore be required to maintain adhesion between epithelial cell layers. If true, $Alr2^{wt/null}$ colonies might be unfit, and $Alr2^{null/null}$ animals might be inviable. This would also place an upper limit on the total frequency of null alleles in a population. A third possibility is that our assay is unable to detect null alleles. This would be the case if cells aggregate in our assay at a lower affinity than that required for colonies to recognize a tissue as self.

Assuming our assay does correlate with the *in vivo* function of Alr2, the observation that three sequences (Anc, 046B, and Hap074) encode the same binding specificity might also seem surprising, since their common specificity would make them less fit than 111A06 or 214E06. *Alr2* allele frequencies might differ from expected equilibrium frequencies if the population has experienced changes in gene flow or recent bottlenecks. Similarly, if there are beneficial alleles at nonallorecognition genes tightly linked to *Alr2*, some *Alr2* alleles might have higher than expected frequencies due to genetic hitchhiking. In addition, we note that the tree for this clade does not represent actual allele frequencies because we removed duplicate sequences prior to constructing the phylogeny. Indeed, in the original study (Gloria-Soria et al., 2012), which reported near-saturation sampling of a single population in Long Island Sound, USA, the 111A06 specificity was represented by three alleles, the Anc/046B/Hap074 specificity by five alleles, and the 214E06 specificity by three alleles. Although essentially anecdotal, this distribution is closer to the expectation of equal phenotype frequencies. These considerations suggest that future work elucidating the population genetics of *Hydractinia*, comprehensively assessing the full breadth of *Alr2* binding specificity diversity, and annotating genomic regions linked to *Alr1* and *Alr2* will be fruitful.

Many positions in domain 1 appear to have experienced positive selection that was either episodic (i.e., limited to particular branches and detected by MEME) or pervasive (under pressure throughout the phylogeny and detected by FEL). As previously mentioned, our evolutionary analyses cannot not tell us whether the six specific mutations that occurred within the branches of our clade experienced positive selection. What we can say is that four mutations occurred at positions under positive selection somewhere in the phylogeny. Two of these mutations (at positions 32 and 44) altered binding specificity and two did not (89 and 93). One interpretation of this is that nonsynonymous mutations are favored at these positions

because they can alter specificity in some sequence contexts, some of which are present in other branches of the tree. Alternatively, these latter mutations might actually alter specificity at a level that our assays could not detect. With respect to position 47, which was not found to be under positive selection but which did alter binding specificity, it is possible that positive selection was present but neither MEME nor FEL had power to detect positive selection because the branches were short. The same explanation could apply to position 76, although our results suggest that positive selection acted only briefly and against a background of strong negative selection. This would also be consistent with mutations at this position altering specificity elsewhere in the larger tree. Several sites in the hypothesized binding surface were also predicted be under negative selection. These sites could be highly conserved because altering them would render the domain incapable of homophilic binding at all. Further work to complement these analyses with functional assays will answer these questions.

Because our study identified residues that affect homophilic binding specificity, we attempted to use structural modeling to identify the biophysical basis of this specificity. Ultimately, we determined that the models produced by I-TASSER were not reliable enough for us to do so. Understanding the biophysical mechanism of this exquisite specificity must therefore await experimentally determined structures. We were, however, able to use sequence variation to generate a hypothesis for *how* the proteins interact. Across all *Alr2* alleles, positions with the highest degree of variation, and those experiencing diversifying selection, were predicted to occur on one side of the Ig-like beta barrel. This suggests that the N-terminal domains of Alr2 bind in a side-to-side manner.

Although we focused here on domain 1, other regions might also determine binding specificity. Evidence for this comes from the fact that the entire extracellular region of Alr2 is polymorphic, and the prediction that residues in domains 2–3 and the ECS are predicted to have experienced diversifying selection (Nicotra et al., 2009; Gloria-Soria et al., 2012). Point mutations in these regions might also give rise to new alleles. Recombination might also generate novel binding specificities. Domains 1–3 and the ECS are each encoded by single exons. These exons frequently recombine between *Alr2* alleles and get shuffled between *Alr2* and several adjacent pseudogenes via gene conversion or unequal crossing over (Gloria-Soria et al., 2012). This could generate chimeric domain 1 sequences with novel specificities. It might also bring together new combinations of domains 1–3 or the ECS that would have different specificities than either of the nonrecombinant parental alleles.

In light of our results, *Hydractinia* would be a productive system in which to study "sequence space"—the theoretical universe of all possible peptides of a given length. Long-standing questions about how many functional variants of a protein exist in sequence space, how many of these actually appear in nature, and whether evolution is constrained in its ability to reach them remain unresolved (Weinreich et al., 2006; Povolotskaya and Kondrashov, 2010; Podgornaia and Laub, 2015). Because natural selection drives the continued evolution of new allorecognition alleles, allorecognition loci like *Alr2* are essentially natural experiments exploring sequence space.

Limitations of the study

The main limitation of this study is the qualitative nature of our cell aggregation assays. Although such assays are commonly used to test binding in cell adhesion molecules (Kasinrerk et al., 1999; Katsamba et al., 2009; Schreiner and Weiner, 2010; Thu et al., 2014; Rubinstein et al., 2015; Goodman et al., 2016), it can be difficult to draw conclusions from them about quantitative binding affinities. This is particularly true here because we used transient transfections, which led to unavoidable variation in the expression of each Alr2 isoform between cell populations. This prevented us from using measures of aggregation speed or aggregate size to infer their binding strength. In other words, in this study, assays with just one allele reveal whether the encoded protein can bind to itself in *trans* but do not indicate its homophilic binding affinity. Similarly, assays in which two alleles are present only reveal whether homophilic or heterophilic interactions were favored. Therefore, it is possible that isoforms that did not bind each other in our assays would, in fact, bind heterophilically if homophilic interactions were prevented, as would likely be the case if they were expressed on the outward facing epithelia of opposing *Hydractinia* colonies. With this limitation in mind, we conservatively interpreted "semimixed" aggregates as indicating that two isoforms had heterophilic affinities that were relatively weaker than their homophilic affinities. We hypothesize this type of aggregate formed because the difference in affinities led to homophilic clusters that then associated heterophilically. This interpretation is in line with what is thought to happen when similar aggregates form with cadherins

and other immunoglobulin superfamily cell adhesion proteins (Katsamba et al., 2009; Goodman et al., 2016). These caveats should be kept in mind when extrapolating our results to nature. Resolving this issue will require quantitative assays paired with transgenic experiments to ectopically express these alleles in living colonies and determine their phenotypic effect, an experimental approach now possible thanks to recent advances in *Hydractinia* functional genomics (Sanders et al., 2018).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Alr2 sequence acquisition and processing
  - Phylogenetic analysis and ancestral state reconstruction
  - Constructs for ectopic expression of Alr2 alleles
  - Expression of *Alr2* alleles in mammalian cells
  - Aggregation assay
  - Sequence variability and visualization of domain 1
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102811.

AUTHOR CONTRIBUTIONS

A.H. and T.C. performed the experiments. A.H. did the data analysis and structural visualizations. A.H. and M.N. designed the experiments and wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Aanen, D.K., Debets, A.J.M., Visser, J.A.G.M. De, and Hoekstra, R.F. (2008). The social evolution of somatic fusion. Bioessays *30*, 1193–1203. https://doi.org/10.1002/bies.20840.

Abràmoff, M.D., Magalhães, P.J., and Ram, S.J. (2004). Image processing with ImageJ. Biophotonics Int. *7*, 36–42.

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

Cadavid, L.F., Powell, A.E., Nicotra, M.L., Moreno, M., and Buss, L.W. (2004). An invertebrate histocompatibility complex. Genetics *167*, 357–365. https://doi.org/10.1534/genetics.167.1.357.

Cao, P., Wei, X., Awal, P., and Müller, R. (2019). A highly polymorphic receptor governs many distinct self-recognition types within the Myxococcales order. MBio *10*, 1–15.

Casselton, L.A., and Olesnicky, N.S. (1998). Molecular genetics of mating recognition in Basidiomycete fungi. Microbiol. Mol. Biol. Rev. *62*, 55–70.

Dubuc, T.Q., Schnitzler, C.E., Chrysostomou, E., Mcmahon, E.T., Gahan, J.M., Buggie, T., Gornik, S.G., Hanley, S., Barreira, S.N., Gonzalez, P., et al. (2020). Transcription factor AP2 controls cnidarian germ cell induction. Science *367*, 757–762.

Dutheil, J., and Boussau, B. (2008). Of Libraries and Programs 12, pp. 1–12. https://doi.org/10.1186/1471-2148-8-255.

Frank, U., Nicotra, M.L., and Schnitzler, C.E. (2020). The colonial cnidarian Hydractinia. Evodevo *11*, 7–12. https://doi.org/10.1186/s13227-020-00151-0.

Fujii, S., Kubo, K., and Takayama, S. (2016). Non-self- and self-recognition models in plant self-incompatibility. Nat. Plants *2*, 1–9. https://doi.org/10.1038/NPLANTS.2016.130.

Gibbs, K.A., and Greenberg, E.P. (2011). Territoriality in proteus: advertisement and aggression. Chem. Rev. *111*, 188–194.

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. https://doi.org/10.1093/molbev/mss197.

Gonçalves, A.P., Heller, J., Rico-ramírez, A.M., Daskalov, A., Rosenfield, G., and Glass, N.L. (2020). conflict, competition, and cooperation regulate social interactions in Filamentous fungi. Annu. Rev. Microbiol. *74*, 693–712.

Goncalves, A.P., Heller, J., Span, E.A., Rosenfiled, G., Do, H.P., Palma-Guerrero, J., Requena, N., Marletta, M.A., and Glass, N.L. (2019). Allorecognition upon fungal cell-cell contact determines social cooperation and impacts the acquisition of multicellularity article allorecognition upon fungal cell-cell contact determines social cooperation. Curr. Biol. *29*, 3006–3017. https://doi.org/10.1016/j.cub.2019.07.060.

Goodman, K.M., Yamagata, M., Jin, X., Mannepalli, S., Katsamba, P.S., Ahlsen, G., Sergeeva, A.P., Honig, B., Sanes, J.R., and Shapiro, L. (2016). Molecular basis of sidekick-mediated cell-cell adhesion and specificity. Elife *5*, 1–21. https://doi.org/10.7554/eLife.19058.

James, T.Y. (2015). Why mushrooms have evolved to be so promiscuous: insights from evolutionary and ecological patterns. Fungal Biol. Rev. *29*, 167–178. https://doi.org/10.1016/j.fbr.2015.10.002.

Karadge, U.B., Gosto, M., and Nicotra, M.L. (2015). Allorecognition proteins in an invertebrate exhibit homophilic interactions. Curr. Biol. *25*, 2845–2850. https://doi.org/10.1016/j.cub.2015.09.030.

Kasinrerk, W., Tokrasinwit, N., and Phunpae, P. (1999). CD147 monoclonal antibodies induce homotypic cell aggregation of monocytic cell line U937 via LFA-1/ICAM-1 pathway. Immunology *96*, 184–192. https://doi.org/10.1046/j.1365-2567.1999.00653.x.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. *33*, 511–518. https://doi.org/10.1093/nar/gki198.

Katsamba, P., Carroll, K., Ahlsen, G., Bahna, F., Vendome, J., Posy, S., Rajebhosale, M., Price, S., Jessell, T.M., Ben-Shaul, A., et al. (2009). Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. Proc. Natl. Acad. Sci. U. S. A. *106*, 11594–11599. https://doi.org/10.1073/pnas.0905349106.

Kimura, M., and Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. Genetics *49*, 725–738.

Kundert, P., and Shaulsky, G. (2019). Cellular allorecognition and its roles in Dictyostelium development and social evolution. Int. J. Dev. Biol. *393*, 383–393.

Künzel, T., Heiermann, R., Frank, U., Müller, W., Tilmann, W., Bause, M., Nonn, A., Helling, M., Schwarz, R.S., and Plickert, G. (2010). Migration and differentiation potential of stem cells in the cnidarian Hydractinia analysed in eGFP-transgenic animals and chimeras. Dev. Biol. *348*, 120–129. https://doi.org/10.1016/j.ydbio.2010.08.017.

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. https://doi.org/10.1016/j.cell.2005.10.026.

Lam, A.J., St-pierre, F., Gong, Y., Marshall, J.D., Cranfill, P.J., Baird, M.A., Mckeown, M.R., Wiedenmann, J., Davidson, M.W., Schnitzer, M.J., et al. (2012). Improving FRET dynamic range with bright green and red fluorescent proteins. Nat. Methods *9*, 1005–1012. https://doi.org/10.1038/NMETH.2171.

Lawrence, M.J. (2000). Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. Ann. Bot. *85*, 221–226.

Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: recent updates and. Nucleic Acids Res. *47*, 256–259. https://doi.org/10.1093/nar/gkz239.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol. Biol. *1079*, 155—170. https://doi.org/10.1007/978-1-62703-646-7_10.

Meng, E.C., Pettersen, E.F., Couch, G.S., Huang, C.C., and Ferrin, T.E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics *7*, 1–10. https://doi.org/10.1186/1471-2105-7-339.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Pond, S.L.K. (2012). Detecting Individual sites subject to episodic diversifying selection. PLoS Genet. *8*, 1–10. https://doi.org/10.1371/journal.pgen.1002764.

Nicotra, M.L. (2019). Invertebrate allorecognition. Curr. Biol. *29*, R463–R467. https://doi.org/10.1016/j.cub.2019.03.039.

Nicotra, M.L., and Buss, L.W. (2005). A test for larval kin aggregations. Biol. Bull. *208*, 157–158.

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. https://doi.org/10.1016/j.cub.2009.02.040.

Nydam, M.L., Stephenson, E.E., Waldman, C.E., and Tomaso, A.W. De (2017). Balancing selection on allorecognition genes in the colonial ascidian Botryllus schlosseri. Dev. Comp. Immunol. *69*, 60–74. https://doi.org/10.1016/j.dci.2016.12.006.

Paoletti, M. (2016). Vegetative incompatibility in fungi: from recognition to cell death, whatever does the trick. Fungal Biol. Rev. *30*, 152–162. https://doi.org/10.1016/j.fbr.2016.08.002.

Pathak, D.T., Wei, X., Dey, A., and Wall, D. (2013). Molecular recognition by a polymorphic cell surface receptor governs cooperative behaviors in bacteria. PLoS Genet. *9*, 1–12. https://doi.org/10.1371/journal.pgen.1003891.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera — a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. https://doi.org/10.1002/jcc.20084.

Podgornaia, A.I., and Laub, M.T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673–677. https://doi.org/10.1126/science.1257360.

Pond, S.L.K., and Frost, S.D.W. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. *22*, 1208–1222. https://doi.org/10.1093/molbev/msi105.

Pond, S.L.K., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., et al. (2019). HyPhy 2.5 — a Customizable platform for evolutionary hypothesis testing using Phylogenies. Mol. Biol. Evol. *37*, 295–299. https://doi.org/10.1093/molbev/msz197.

Povolotskaya, I.S., and Kondrashov, F.A. (2010). Sequence space and the ongoing expansion of the protein universe. Nature *465*, 922–926. https://doi.org/10.1038/nature09105.

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. G *1*, 499–503. https://doi.org/10.1534/g3.111.001149.

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. https://doi.org/10.1534/genetics.107.075689.

Richman, A.D., and Kohn, J.R. (2000). Evolutionary genetics of self-incompatibility in the Solanaceae. Plant Mol. Biol. *42*, 169–179.

Rosa, S.F.P., 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. https://doi.org/10.1016/j.cub.2010.04.050.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738. https://doi.org/10.1038/nprot.2010.5.

Rubinstein, R., Thu, C.A., Goodman, K.M., Wolcott, H.N., Bahna, F., Mannepalli, S., Ahlsen, G., Chevee, M., Halim, A., Clausen, H., et al. (2015). Molecular logic of neuronal self-recognition through protocadherin domain interactions. Cell *163*, 629–642. https://doi.org/10.1016/j.cell.2015.09.026.

Sanders, S.M., Ma, Z., Hughes, J.M., Riscoe, B.M., Gibson, G.A., Watson, A.M., Flici, H., Frank, U., Schnitzler, C.E., Baxevanis, A.D., and Nicotra, M.L. (2018). CRISPR/Cas9-mediated gene knockin in the hydroid Hydractinia symbiolongicarpus. BMC Genomics 19, 1–17. https://doi.org/10.1186/s12864-018-5032-z.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675. https://doi.org/10.1038/nmeth.2089.

Schreiner, D., and Weiner, J.a (2010). Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. Proc. Natl. Acad. Sci. U. S. A. 107, 14893–14898. https://doi.org/10.1073/pnas.1004526107.

Spielman, S.J., Weaver, S., Shank, S.D., Magalis, B.R., Li, M., and Pond, S.L.K. (2019). Evolution of viral genomes: interplay between selection, recombination, and other forces. In Evolutionary Genomics. Methods in Molecular Biology, pp. 427–468. https://doi.org/10.1007/978-1-4939-9074-0.

Stoner, D.S., Rinkevich, B., and Weissman, I.L. (1999). Heritable germ and somatic cell lineage competitions in chimeric colonial protochordates. Proc 96, 9148–9153.

Stoner, D.S., and Weissman, I.L. (1996). Somatic and germ cell parasitism in a colonial ascidian: possible role for a highly polymorphic allorecognition system. Proc. Natl. Acad. Sci. U. S. A. 93, 15254–15259.

Thu, C.A., Chen, W.V., Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between alpha, beta,and gamma protocadherins. Cell 158, 1045–1059. https://doi.org/10.1016/j.cell.2014.07.012.

Trifinopoulos, J., Nguyen, L., Haeseler, A.V., and Minh, B.Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 44, 232–235. https://doi.org/10.1093/nar/gkw256.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview

Version 2 — a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

Weinreich, D.M., Delaney, N.F., DePristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312, 111–114. https://doi.org/10.1126/science.1123539.

Wright, S. (1939). The distribution of self-sterility alleles in populations. Genetics 24, 538–552.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER suite: protein structure and function prediction. Nat. Publ. Gr. 12, 7–8. https://doi.org/10.1038/nmeth.3213.

Yund, P., Cunningham, C.W., and Buss, L.W. (1987). Recruitment and postrecruitment interactions in a colonial hydroid. Ecology 68, 971–982. https://doi.org/10.2307/1938368.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9, 1–18. https://doi.org/10.1186/1471-2105-9-40.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: cell lines		
HEK293T cells	ATCC	Cat# CRL-3216
Recombinant DNA		
pFLAG-CMV3-111A06-eGFP	This paper	pUP801
pFLAG-CMV3-111A06-mRuby2	This paper	pUP738
pFLAG-CMV3-214E06-eGFP	This paper	pUP836
pFLAG-CMV3-214E06-mRuby2	This paper	pUP746
pFLAG-CMV3-Anc-eGFP	This paper	pUP871
pFLAG-CMV3-Anc-mRuby2	This paper	pUP748
pFLAG-CMV3-046B-eGFP	This paper	pUP872
pFLAG-CMV3-046B-mRuby2	This paper	pUP750
pFLAG-CMV3-Hap074-eGFP	This paper	pUP894
pFLAG-CMV3-Hap074-mRuby2	This paper	pUP752
pFLAG-CMV3-214E06-N32Y-eGFP	This paper	pUP838
pFLAG-CMV3-214E06-N32Y-mRuby2	This paper	pUP839
pFLAG-CMV3-Hap074-G47E-eGFP	This paper	pUP875
pFLAG-CMV3-Hap074-G47E-mRuby2	This paper	pUP876
pFLAG-CMV3-Hap074-S44G-eGFP	This paper	pUP878
pFLAG-CMV3-Hap074-S44G-mRuby2	This paper	pUP877
pFLAG-CMV3-Anc-T76R-eGFP	This paper	pUP866
pFLAG-CMV3-Anc-T76R-mRuby2	This paper	pUP865
pFLAG-CMV3-Anc-E93K-eGFP	This paper	pUP879
pFLAG-CMV3-Anc-E93K-mRuby2	This paper	pUP880
pFLAG-CMV3-046B-N32Y-eGFP	This paper	pUP892
pFLAG-CMV3-046B-N32Y-mRuby2	This paper	pUP794
pFLAG-CMV3-Hap074-N32Y-eGFP	This paper	pUP893
pFLAG-CMV3-Hap074-N32Y-mRuby2	This paper	pUP795
Software and algorithms		
I-TASSER v5.1	Roy et al. (2010); Yang et al. (2015); Zhang (2008)	https://zhanglab.ccmb.med.umich.edu/I-TASSER/download/
Multialign Viewer (Chimera v1.14)	USCF Chimera Meng et al., 2006; Pettersen et al. (2004)	NA
MEME (HyPhy 2.2.4)	Murrell et al., 2012	http://hyphy.org/w/index.php/Download
FEL (HyPhy 2.2.4)	Pond and Frost (2005)	http://hyphy.org/w/index.php/Download
ImageJ v1.53a	Abràmoff et al., 2004; Schneider et al. (2012)	https://imagej.nih.gov/ij/
Other		
TransIT-293	Mirus Bio	Cat#MIR 2700
35μm strainer mesh	Steller scientific	Cat#FSC-FLTCP

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
DNase I	Sigma	Cat#D4527-10KU
24-well ultra-low attachment plate	Fisher Scientific	Cat#07-200-602
Orbital rotator	IBI Scientific	Model# BBUAAUVIS
NEBuilder HiFi DNA Assembly	New England Biolabs	Cat#E2621S
Hap200	Genbank	JX048906.1
Hap199	Genbank	JX048905.1
Hap197	Genbank	JX048904.1
Hap196	Genbank	JX048903.1
Hap195	Genbank	JX048902.1
Hap194	Genbank	JX048901.1
Hap193	Genbank	JX048900.1
Hap192	Genbank	JX048899.1
Hap191	Genbank	JX048898.1
Hap190	Genbank	JX048897.1
Hap189	Genbank	JX048896.1
Hap188	Genbank	JX048895.1
Hap187	Genbank	JX048894.1
Hap186	Genbank	JX048893.1
Hap185	Genbank	JX048892.1
Hap184	Genbank	JX048891.1
Hap183	Genbank	JX048890.1
Hap182	Genbank	JX048889.1
Hap181	Genbank	JX048888.1
Hap180	Genbank	JX048887.1
Hap179	Genbank	JX048886.1
Hap178	Genbank	JX048885.1
Hap176	Genbank	JX048884.1
Hap175	Genbank	JX048883.1
Hap173	Genbank	JX048881.1
Hap172	Genbank	JX048880.1
Hap171	Genbank	JX048879.1
Hap169	Genbank	JX048877.1
Hap168	Genbank	JX048876.1
Hap167	Genbank	JX048875.1
Hap166	Genbank	JX048874.1
Hap165	Genbank	JX048873.1
Hap164	Genbank	JX048872.1
Hap163	Genbank	JX048871.1
Hap161	Genbank	JX048869.1
Hap159	Genbank	JX048867.1
Hap158	Genbank	JX048866.1
Hap157	Genbank	JX048865.1
Hap156	Genbank	JX048864.1
Hap155	Genbank	JX048863.1
Hap154	Genbank	JX048862.1
Hap153	Genbank	JX048861.1

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Hap152	Genbank	JX048860.1
Hap151	Genbank	JX048859.1
Hap150	Genbank	JX048858.1
Hap149	Genbank	JX048857.1
Hap148	Genbank	JX048856.1
Hap147	Genbank	JX048855.1
Hap146	Genbank	JX048854.1
Hap145	Genbank	JX048853.1
Hap144	Genbank	JX048852.1
Hap143	Genbank	JX048851.1
Hap142	Genbank	JX048850.1
Hap141	Genbank	JX048849.1
Hap139	Genbank	JX048847.1
Hap138	Genbank	JX048846.1
Hap137	Genbank	JX048845.1
Hap136	Genbank	JX048844.1
Hap135	Genbank	JX048843.1
Hap134	Genbank	JX048842.1
Hap133	Genbank	JX048841.1
Hap132	Genbank	JX048840.1
Hap131	Genbank	JX048839.1
Hap130	Genbank	JX048838.1
Hap129	Genbank	JX048837.1
Hap128	Genbank	JX048836.1
Hap127	Genbank	JX048835.1
Hap126	Genbank	JX048834.1
Hap125	Genbank	JX048833.1
Hap124	Genbank	JX048832.1
Hap123	Genbank	JX048831.1
Hap122	Genbank	JX048830.1
Hap121	Genbank	JX048829.1
Hap120	Genbank	JX048828.1
Hap119	Genbank	JX048827.1
Hap118	Genbank	JX048826.1
Hap117	Genbank	JX048825.1
Hap116	Genbank	JX048824.1
Hap115	Genbank	JX048823.1
Hap114	Genbank	JX048822.1
Hap113	Genbank	JX048821.1
Hap112	Genbank	JX048820.1
Hap111	Genbank	JX048819.1
Hap110	Genbank	JX048818.1
Hap109	Genbank	JX048817.1
Hap108	Genbank	JX048816.1
Hap107	Genbank	JX048815.1
Hap106	Genbank	JX048814.1

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Hap105	Genbank	JX048813.1
Hap104	Genbank	JX048812.1
Hap103	Genbank	JX048811.1
Hap102	Genbank	JX048810.1
Hap101	Genbank	JX048809.1
Hap100	Genbank	JX048808.1
Hap099	Genbank	JX048807.1
Hap098	Genbank	JX048806.1
Hap097	Genbank	JX048805.1
Hap096	Genbank	JX048804.1
Hap095	Genbank	JX048803.1
Hap094	Genbank	JX048802.1
Hap093	Genbank	JX048801.1
Hap092	Genbank	JX048800.1
Hap091	Genbank	JX048799.1
Hap090	Genbank	JX048798.1
Hap089	Genbank	JX048797.1
Hap088	Genbank	JX048796.1
Hap087	Genbank	JX048795.1
Hap086	Genbank	JX048794.1
Hap085	Genbank	JX048793.1
Hap084	Genbank	JX048792.1
Hap083	Genbank	JX048791.1
Hap082	Genbank	JX048790.1
Hap081	Genbank	JX048789.1
Hap080	Genbank	JX048788.1
Hap079	Genbank	JX048787.1
Hap078	Genbank	JX048786.1
Hap077	Genbank	JX048785.1
Hap076	Genbank	JX048784.1
Hap075	Genbank	JX048783.1
Hap074	Genbank	JX048782.1
Hap073	Genbank	JX048781.1
Hap072	Genbank	JX048780.1
Hap071	Genbank	JX048779.1
Hap070	Genbank	JX048778.1
Hap069	Genbank	JX048777.1
Hap068	Genbank	JX048776.1
Hap067	Genbank	JX048775.1
Hap066	Genbank	JX048774.1
Hap065	Genbank	JX048773.1
Hap064	Genbank	JX048772.1
Hap063	Genbank	JX048771.1
Hap062	Genbank	JX048770.1
Hap061	Genbank	JX048769.1

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Hap060	Genbank	JX048768.1
Hap058	Genbank	JX048766.1
Hap057	Genbank	JX048765.1
Hap056	Genbank	JX048764.1
Hap054	Genbank	JX048762.1
Hap051	Genbank	JX048759.1
Hap048	Genbank	JX048756.1
Hap046	Genbank	JX048754.1
Hap045	Genbank	JX048753.1
Hap044	Genbank	JX048752.1
Hap043	Genbank	JX048751.1
Hap042	Genbank	JX048750.1
Hap041	Genbank	JX048749.1
Hap040	Genbank	JX048748.1
Hap039	Genbank	JX048747.1
Hap038	Genbank	JX048746.1
Hap037	Genbank	JX048745.1
Hap036	Genbank	JX048744.1
Hap035	Genbank	JX048743.1
Hap034	Genbank	JX048742.1
Hap033	Genbank	JX048741.1
Hap032	Genbank	JX048740.1
Hap031	Genbank	JX048739.1
Hap030	Genbank	JX048738.1
Hap029	Genbank	JX048737.1
Hap028	Genbank	JX048736.1
Hap027	Genbank	JX048735.1
Hap026	Genbank	JX048734.1
Hap025	Genbank	JX048733.1
Hap024	Genbank	JX048732.1
Hap022	Genbank	JX048730.1
Hap021	Genbank	JX048729.1
Hap020	Genbank	JX048728.1
Hap019	Genbank	JX048727.1
Hap018	Genbank	JX048726.1
Hap017	Genbank	JX048725.1
Hap016	Genbank	JX048724.1
Hap015	Genbank	JX048723.1
Hap014	Genbank	JX048722.1
Hap012	Genbank	JX048720.1
Hap010	Genbank	JX048718.1
Hap008	Genbank	JX048716.1
Hap007	Genbank	JX048715.1
Hap006	Genbank	JX048714.1
Hap005	Genbank	JX048713.1

CellPress
OPEN ACCESS

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Hap004	Genbank	JX048712.1
Hap003	Genbank	JX048711.1
Hap002	Genbank	JX048710.1
Hap001	Genbank	JX048709.1
Hap174	Genbank	JX048882.1
Hap162	Genbank	JX048870.1
Hap160	Genbank	JX048868.1
Hap050	Genbank	JX048758.1
Hap013	Genbank	JX048721.1
Hap009	Genbank	JX048717.1
LH09_466G04	Genbank	JX049024.1
LH09_466B06	Genbank	JX049023.1
LH09_465F03	Genbank	JX049022.1
LH09_465B09	Genbank	JX049021.1
LH09_459C06	Genbank	JX049020.1
LH09_459C03	Genbank	JX049019.1
LH09_454D03	Genbank	JX049018.1
LH09_452H02	Genbank	JX049017.1
LH09_449H03	Genbank	JX049016.1
LH09_449F06	Genbank	JX049015.1
LH09_447F08	Genbank	JX049014.1
LH09_447D09	Genbank	JX049013.1
LH09_443D04	Genbank	JX049012.1
LH09_443B07	Genbank	JX049011.1
LH09_436B04	Genbank	JX049010.1
LH09_435B06	Genbank	JX049009.1
LH09_435B05	Genbank	JX049008.1
LH09_431F06	Genbank	JX049007.1
LH09_431C08	Genbank	JX049006.1
LH09_429C03	Genbank	JX049005.1
LH09_429A03	Genbank	JX049004.1
LH09_425B08	Genbank	JX049003.1
LH09_425B07	Genbank	JX049002.1
LH09_417C08	Genbank	JX049001.1
LH09_417B05	Genbank	JX049000.1
LH09_406B01	Genbank	JX048999.1
LH09_396C10	Genbank	JX048998.1
LH09_396B06	Genbank	JX048997.1
LH09_396A08	Genbank	JX048996.1
LH09_395G03	Genbank	JX048995.1
LH09_386C08	Genbank	JX048994.1
LH09_384F06	Genbank	JX048993.1
LH09_384E03	Genbank	JX048992.1
LH09_380B02	Genbank	JX048991.1
LH09_380A03	Genbank	JX048990.1

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
LH09_274C02	Genbank	JX048989.1
LH09_271B05	Genbank	JX048988.1
LH09_270D09	Genbank	JX048987.1
LH09_270C02	Genbank	JX048986.1
LH09_268E09	Genbank	JX048985.1
LH09_268B05	Genbank	JX048984.1
LH09_265F08	Genbank	JX048983.1
LH09_265B10	Genbank	JX048982.1
LH09_261C01	Genbank	JX048981.1
LH09_249C04	Genbank	JX048980.1
LH09_249B04	Genbank	JX048979.1
LH09_248A06	Genbank	JX048978.1
LH09_244H05	Genbank	JX048977.1
LH09_244B05	Genbank	JX048976.1
LH09_230G03	Genbank	JX048975.1
LH09_230F04	Genbank	JX048974.1
LH09_230B04	Genbank	JX048973.1
LH09_214H10	Genbank	JX048972.1
LH09_214E06	Genbank	JX048971.1
LH09_212C05	Genbank	JX048970.1
LH09_212B03	Genbank	JX048969.1
LH09_205E03	Genbank	JX048968.1
LH09_205C02	Genbank	JX048967.1
LH09_202E04	Genbank	JX048966.1
LH09_162G06	Genbank	JX048965.1
LH09_162D02	Genbank	JX048964.1
LH09_158D11	Genbank	JX048963.1
LH09_158A08	Genbank	JX048962.1
LH09_158A03	Genbank	JX048961.1
LH09_145D12	Genbank	JX048960.1
LH09_116B02	Genbank	JX048959.1
LH09_111C09	Genbank	JX048958.1
LH09_111A06	Genbank	JX048957.1
LH09_110C01	Genbank	JX048956.1
LH09_110B01	Genbank	JX048955.1
LH09_085A06	Genbank	JX048954.1
LH09_084B07	Genbank	JX048953.1
LH09_083D05	Genbank	JX048952.1
LH09_083C10	Genbank	JX048951.1
LH09_082F03	Genbank	JX048950.1
LH09_082D07	Genbank	JX048949.1
LH09_078E08	Genbank	JX048948.1
LH09_068F07	Genbank	JX048947.1
LH09_068B01	Genbank	JX048946.1
LH09_064D04	Genbank	JX048945.1

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
LH09_064C05	Genbank	JX048944.1
LH09_061G09	Genbank	JX048943.1
LH09_061G06	Genbank	JX048942.1
LH09_059C03	Genbank	JX048941.1
LH09_058G02	Genbank	JX048940.1
LH09_058C05	Genbank	JX048939.1
LH09_055H01	Genbank	JX048938.1
LH09_055A07	Genbank	JX048937.1
LH09_054E03	Genbank	JX048936.1
LH09_054D05	Genbank	JX048935.1
LH09_052D03	Genbank	JX048934.1
LH09_052C02	Genbank	JX048933.1
LH09_051E03	Genbank	JX048932.1
LH09_051B03	Genbank	JX048931.1
LH09_048E02	Genbank	JX048930.1
LH09_044F06	Genbank	JX048929.1
LH09_044C10	Genbank	JX048928.1
LH09_042B02	Genbank	JX048927.1
LH09_042A05	Genbank	JX048926.1
LH09_039G03	Genbank	JX048925.1
LH09_037B08	Genbank	JX048924.1
LH09_037A08	Genbank	JX048923.1
LH09_035C05	Genbank	JX048922.1
LH09_034H09	Genbank	JX048921.1
LH09_032F03	Genbank	JX048920.1
LH09_032E02	Genbank	JX048919.1
LH09_024B01	Genbank	JX048918.1
LH09_023H02	Genbank	JX048917.1
LH09_023E08	Genbank	JX048916.1
LH09_019D10	Genbank	JX048915.1
LH09_018D08	Genbank	JX048914.1
LH09_018B08	Genbank	JX048913.1
LH09_005G06	Genbank	JX048912.1
LH09_005E05	Genbank	JX048911.1
LH09_004E06	Genbank	JX048910.1
LH09_004B06	Genbank	JX048909.1
LH09_001E02	Genbank	JX048908.1
14_7F	Genbank	JX048907.1
OQ-6Db	Genbank	HM013632.1
OQ-6Da	Genbank	HM013631.1
LH07:060b	Genbank	HM013630.1
LH07:060a	Genbank	HM013629.1
LH07:049b	Genbank	HM013628.1
LH07:049a	Genbank	HM013627.1
LH07:046b	Genbank	HM013626.1

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
LH07:046a	Genbank	HM013625.1
LH07:043a	Genbank	HM013624.1
LH07:041b	Genbank	HM013623.1
LH07:041a	Genbank	HM013622.1
LH07:037b	Genbank	HM013621.1
LH07:037a	Genbank	HM013620.1
LH07:036b	Genbank	HM013619.1
LH07:036a	Genbank	HM013618.1
LH07:026b	Genbank	HM013617.1
LH07:026a	Genbank	HM013616.1
LH06:050b	Genbank	HM013613.1
LH06:049b	Genbank	HM013611.1
LH06:028b	Genbank	HM013609.1
LH06:028a	Genbank	HM013608.1
LH06:003b	Genbank	HM013607.1
alr2-W60b	Genbank	FJ207419.1
alr2-W60a	Genbank	FJ207418.1
alr2-W49b	Genbank	FJ207417.1
alr2-W49a	Genbank	FJ207416.1
alr2-W41b	Genbank	FJ207415.1
alr2-W41a	Genbank	FJ207414.1
alr2-W36b	Genbank	FJ207413.1
alr2-W36a	Genbank	FJ207412.1
alr2-W14b	Genbank	FJ207411.1
alr2-W14a	Genbank	FJ207410.1
alr2-LH49b	Genbank	FJ207402.1
alr2-LH49a	Genbank	FJ207401.1
alr2-LH03i	Genbank	FJ207396.1
alr2-LH03a	Genbank	FJ207395.1
LH07:014b	Genbank	HM013615.1
LH07:014a	Genbank	HM013614.1
LH06:050a	Genbank	HM013612.1
LH06:049a	Genbank	HM013610.1
LH06:003a	Genbank	HM013606.1
alr2-LH53b	Genbank	FJ617568.1
alr2-LH53a	Genbank	FJ617567.1
alr2-LH04b	Genbank	FJ617566.1
alr2-LH04a	Genbank	FJ617565.1
alr2-R	Genbank	FJ207409.1
alr2-LH82b	Genbank	FJ207408.1
alr2-LH82a	Genbank	FJ207407.1
alr2-LH58b	Genbank	FJ207406.1
alr2-LH58a	Genbank	FJ207405.1
alr2-LH57b	Genbank	FJ207404.1
alr2-LH57a	Genbank	FJ207403.1

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
alr2-LH09b	Genbank	FJ207400.1
alr2-LH09a	Genbank	FJ207399.1
alr2-LH08b	Genbank	FJ207398.1
alr2-LH08a	Genbank	FJ207397.1

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Matthew Nicotra (matthew.nicotra@pitt.edu).

Materials availability

This study did not generate new reagents. Plasmids generated in this study are available from the Lead Contact upon request.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HEK293T cells (ATCC Cat# CRL-3216) were cultured at 37°C with 5% CO_2 in accordance with ATCC guidelines. Complete HEK culture medium was made using DMEM (Fisher Science, SH30081.01), 10% fetal bovine serum (Thermofisher Scientific, #16000044), 0.001% beta-mercaptoethanol (Fisher Scientific, 21-985-023), 100 U/mL penicillin and 100 mg/mL streptomycin (Sigma, P4333-100ML).

METHOD DETAILS

Alr2 sequence acquisition and processing

Alr2 alleles 111A06 and 214E06 were identified from previously published *Alr2* sequences (Gloria-Soria et al., 2012). To obtain a dataset of *Alr2* domain 1 sequences, we downloaded all 373 *Hydractinia symbiolongicarpus Alr2* cDNA sequences from GenBank, aligned them with MAFFT (Katoh et al., 2005) as implemented in Jalview 2.10.5 (Waterhouse et al., 2009), then trimmed the alignment leaving only the region encoding domain 1. Duplicate sequences were then removed with ElimDupes (www.hiv.lanl.gov), to yield 146 distinct domain 1 cDNA sequences, encoding 137 distinct amino acid sequences.

Phylogenetic analysis and ancestral state reconstruction

The 146 domain 1 cDNA sequences were aligned with PRANK (Löytynoja, 2014), a codon-aware alignment program (File S1). The alignment was then used to construct a phylogenetic tree using maximum likelihood through IQ-TREE (http://iqtree.cibiv.univie.ac.at/) (Trifinopoulos et al., 2016) (File S2). From the web portal, the defaults settings were used with codon selected for the sequence type, standard/universal genetic code, ultrafast bootstrap analysis with a maximum of 1000 alignments, 0.99 minimum correlation coefficient, 1000 replicates of SH-aLRT branch test, 0.5 perturbation strength, and 100 set for the IQ-TREE stopping rule. Ancestral states were estimated using the phylogenetic tree generated from IQ-TREE and the ancestral reconstruction function within PRANK (File S3) (Dutheil and Boussau, 2008; Löytynoja, 2014). An unrooted tree was generated using iTOL v5.5.1 with one iteration of equal-daylight (Letunic and Bork, 2019).

Constructs for ectopic expression of Alr2 alleles

The plasmid backbone used for all constructs in this study was the pFLAG-CMV-3 (Sigma, E6783). Previously, it was determined that the N-terminal FLAG tag did not have an effect on the binding capability of Alr2 (Karadge et al., 2015). The *Hydractinia Alr2* allele sequences were optimized for human expression using the Integrated DNA Technologies (IDT) Codon Optimization Tool (https://www.idtdna.com/

CodonOpt). The full Alr2 sequence (domain 1 in the ectodomain through the cytoplasmic tail) for 111A06 and domain 1 sequences for Anc, 046B, Hap074, and 214×10^6 were ordered as gBlocks Gene Fragments from IDT. All other mutant domain sequences were ordered from Twist Bioscience as Gene Fragments. Coding sequences for fluorescent proteins were cloned from vectors encoding eGFP and mRuby2 (gift from Michael Davidson, Addgene plasmid #54614 (Lam et al., 2012)). Cloning was performed using the NEBuilder HiFi DNA Assembly (New England Biolabs, E2621S) with primers designed to amplify the vector and insert sequences with ≥20 bp overlap. The FLAG-111A06-eGFP/mRuby2 plasmids (pUP801, pUP746) were cloned first and then used as the template for cloning in the other domain 1 isoforms. Within the construct, linker sequences were used before (Leu-Ala-Ala-Ala) and after (Gly-Pro-Pro-Val-Glu-Lys) the Alr2 allele.

Expression of *Alr2* alleles in mammalian cells

To prepare plasmids for transfection, plasmids were transformed into chemically competent bacteria and isolated from cultures using the GeneJET Plasmid Midi-prep Kit (Thermofisher Scientific, K0481) or the PureLink HiPure Plasmid Maxiprep Kit (Thermofisher Scientific, K2100006). Plasmids were transiently transfected into HEK293T cells using TransIT-293 (Mirus Bio, MIR 2700) according to the manufacturer's instructions. To summarize, on day 1, HEK293T cells were plated in a 12-well plate (Fisher Scientific, #353043) at a density of 3×10^5/well in 1 mL of complete HEK medium to achieve approximately 60-70% confluency on Day 2. On Day 2, the transfection mixture was prepared in a total volume of 100 μL using 1 ug (X μl) of plasmid DNA (plasmid concentrations between 300ng and 1000ng/μl), diluted with optiMEM (Gibco, #31985-070) (97-X μl), and 3 μL of TransIT-293 reagent. While incubating the DNA:lipid complexes, the cells were washed using 500 μL of DPBS (Fisher Scientific, BW17-512F), incubated with 1 mL transfection medium (complete HEK medium without antibiotics), and replaced in the 5% CO_2 incubator. Once the DNA:lipid complexes had incubated, the 100 μL mixture was added to the appropriate well, the plate gently shaken back and forth and then replaced in the incubator. On Day 4, cells were used in the aggregation assay.

Aggregation assay

Our aggregation protocol is adapted from previous work (Karadge et al., 2015). To summarize, previously transfected HEK293T cells were incubated with 0.25% Trypsin/0.1% EDTA solution (Corning, MT25053CI), washed in complete HEK culture medium, mechanically disrupted via pipette, and filtered through a 35μm strainer mesh (Steller scientific, FSC-FLTCP) to create a single cell suspension. For each aggregation assay, a total of 5×10^4 cells were resuspended in 500 μL aggregation assay medium (complete HEK medium, 70 U/ml DNase I [Sigma, D4527-10KU], and 2 mM EGTA [Goldbio, E−217-25]) and added to one well of a 24-well ultra-low attachment plate (Fisher Scientific, 07-200-602). When testing isoforms pairwise, 2.5×10^4 cells of each transfection were added to the same well and resuspended in a total of 500 μL. The plate was incubated for one hour at 37°C in 5% CO_2 on an orbital rotator (IBI Scientific, Model# BBUAAUVIS) set at 90 rpm. Assays were visualized using an inverted fluorescence microscope (Nikon Eclipse TS100). Each pairwise assay was repeated at least three times. In cases when the assay results could not be viewed immediately, cell aggregates were fixed by adding 500 μL of 8% paraformaldehyde (Fisher, AA433689M) diluted in DPBS to each well and the results imaged within 5 h. All images and merged images were processed using ImageJ (Abràmoff et al., 2004; Schneider et al., 2012).

Sequence variability and visualization of domain 1

The structure for the Anc domain 1 isoform was predicted using I-TASSER v5.1 (Zhang, 2008; Roy et al., 2010; Yang et al., 2015) which resulted in a domain with a V-set like fold. To visualize the variable positions within domain 1, the aligned 137 protein sequences were uploaded to the Multialign Viewer in UCSF Chimera (Pettersen et al., 2004; Meng et al., 2006) and the conservation rendered onto the structure. Sites under positive selection were identified using MEME (Murrell et al., 2012) and FEL (Pond and Frost, 2005) as implemented in HyPhy 2.5.8 (Pond et al., 2019). Both algorithms were run using synonymous rate variation and significance threshold of p = 0.1, as recommended by the developers (Spielman et al., 2019).

QUANTIFICATION AND STATISTICAL ANALYSIS

Sites under positive and/or negative selection were identified using statistical tests as implemented in MEME and FEL, at significance thresholds of p = 0.1. No other quantification or statistical analyses were performed in this study.