

**Molecular basis of allorecognition in the colonial cnidarian *Hydractinia symbiolongicarpus***

by

**Aidan Lorraine Huene**

B.S., Oral Roberts University, 2015

Submitted to the Graduate Faculty of  
School of Medicine in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH  
SCHOOL OF MEDICINE

This dissertation was presented

by

**Aidan Lorraine Huene**

It was defended on

December 3, 2021

and approved by

Andrew VanDemark, Associate Professor, Department of Biological Sciences

Anne-Ruxandra Carvunis, Assistant Professor, Department of Computational and Systems  
Biology

Jon Boyle, Associate Professor, Department of Biological Sciences

Adam Kwiatkowski, Associate Professor, Department of Cell Biology

Dissertation Director: Matthew Nicotra, Assistant Professor, Departments of Surgery and  
Immunology

Copyright © by Aidan Lorraine Huene

2021

# **Molecular basis of allorecognition in the colonial cnidarian *Hydractinia symbiolongicarpus***

Aidan Lorraine Huene, PhD

University of Pittsburgh, 2021

Allorecognition is the ability to distinguish between self tissues and those of conspecifics. Among cnidarians, the molecular basis of allorecognition is best understood in the colonial hydroid, *Hydractinia*. Previous work has established that allorecognition is controlled by at least two linked genes, *Alr1* and *Alr2*, which are located in a genomic region called the Allorecognition Complex (ARC). Both genes encode transmembrane proteins that have two or three extracellular domains. Both *Alr1* and *Alr2* are highly polymorphic and homophilically bind in *trans* in an isoform-specific manner. This had led to the hypothesis that *trans* interactions between matching *Alr1* and *Alr2* isoforms are involved in determining recognition specificity in *Hydractinia*. The evolutionary history of *Alr1* and *Alr2* is unclear and no homologs have been identified to date. In addition, how new alleles evolve which bear unique binding specificities is unclear due to a lack of closely related isoforms to test. Recently, advancements in sequencing have made it feasible to sequence the *Hydractinia* genome. I assembled and annotated three ARC haplotypes that were generated from genomic sequence obtained from a homozygous inbred *Hydractinia* (ARC-F) and a heterozygous wildtype *Hydractinia* (ARC-wt). Here, I report the identification of 41 *Alr*-like loci in ARC-F and 56 *Alr*-like loci in ARC-wt which all share a similar domain architecture. I show that *Alr* sequences encode domains that are novel members of the Immunoglobulin Superfamily predicted to have V-set, I-set, or Fibronectin III-like topologies. The *Alr* loci encode significantly different sequences and in most cases do not align. Comparing multiple alleles of the same *Alr* revealed that most *Alr* alleles appear to be very similar, though not identical. In addition to *Alr1*

and *Alr2*, a third highly polymorphic gene, *Alr6*, was identified. I also report my work showing that the homophilic binding specificity in *Alr2* can evolve rapidly. I showcase this through allelic isoforms with six single amino acid changes that have three distinct binding specificities. These results show that allorecognition in *Hydractinia* is a highly specific, complex mechanism and remains one of the best described recognition systems observed in invertebrates.

## Table of Contents

<b>Acknowledgements .....</b>	<b>xviii</b>
<b>1.0 Introduction.....</b>	<b>1</b>
<b>1.1 Self and non-self recognition in nature.....</b>	<b>1</b>
<b>1.2 Bacterial allorecognition .....</b>	<b>2</b>
1.2.1 <i>Myxococcus xanthus</i> .....	4
1.2.2 <i>Proteus mirabilis</i> .....	6
<b>1.3 Fungal allorecognition.....</b>	<b>8</b>
1.3.1 <i>Neurospora crassa</i> .....	10
1.3.2 <i>Schizophyllum commune</i> .....	12
<b>1.4 Plant allorecognition .....</b>	<b>14</b>
1.4.1 Brassicaceae .....	15
1.4.2 Solanaceae.....	16
1.4.3 Papaveraceae .....	18
1.4.4 Evolution and selection of S-determinant alleles in plant models .....	20
<b>1.5 Slime mold allorecognition .....</b>	<b>21</b>
<b>1.6 Invertebrate allorecognition .....</b>	<b>24</b>
1.6.1 <i>Botryllus schlosseri</i> .....	25
1.6.2 <i>Amphimedon queenslandica</i> .....	27
1.6.3 <i>Hydractinia symbiolongicarpus</i> .....	29
<b>1.7 Common themes in allorecognition .....</b>	<b>34</b>

## 2.0 Sequencing and annotation of the Allorecognition Complex yields a 12 Mb region

which contains a large Alr family .....	37
2.1 Foreword .....	37
2.2 Summary .....	37
2.3 Introduction .....	38
2.4 Results.....	39
2.4.1 Assembly of ARC reference sequence .....	39
2.4.2 The <i>Hydractinia</i> genome contains a large family of Alr genes and pseudogenes .....	43
2.4.3 Alternative splicing alters the domain architecture of several Alr gene products.....	48
2.4.4 Sequences of Alr family members are highly diverse .....	50
2.5 Discussion .....	53
2.6 Methods .....	56
2.6.1 Sequencing and assembly of the genome of an ARC homozygous animal ...	56
2.6.2 RNA extraction and sequencing .....	57
2.6.3 Assembly of the ARC .....	58
2.6.4 Annotation of <i>Alr</i> genes .....	59
3.0 The <i>Alr</i> gene family encodes domains that are novel members of IgSF proteins .....	60
3.1 Foreword .....	60
3.2 Summary .....	60
3.3 Introduction .....	61
3.4 Results.....	63

3.4.1 Domain 1 is an Ig domain most similar to the V-set family. ....	63
3.4.2 Domains 2 and 3 are IgSF domains most similar to the I-set family.....	75
3.4.3 Part of the ECS adopts an immunoglobulin-like fold.....	85
3.4.4 The membrane proximal domains of Alr proteins have six conserved cysteines.....	93
3.4.5 The cytoplasmic tails of many Alr proteins contain ITAM or ITIM motifs	96
3.5 Discussion .....	99
3.6 Methods .....	103
3.6.1 Protein sequence analysis .....	103
3.6.2 Alr sequence comparisons .....	103
3.6.3 Structural predictions and visualization.....	104
<b>4.0 New self-identities evolve via point mutation in an invertebrate allorecognition gene.....</b>	<b>105</b>
4.1 Foreword .....	105
4.2 Summary .....	105
4.3 Introduction .....	106
4.4 Results.....	108
4.4.1 Point mutations in domain 1 can create new binding specificities .....	108
4.4.2 New homophilic specificities can evolve via less restricted intermediates..	113
4.4.3 The N32Y mutation preserves homophilic binding and alters specificity ..	118
4.4.4 Structural and evolutionary analyses suggest a potential binding interface .....	120
4.5 Discussion .....	129



<b>4.6 Methods .....</b>	<b>134</b>
<b>4.6.1 Experimental model and subject details .....</b>	<b>134</b>
<b>4.6.2 Alr2 sequence acquisition and processing .....</b>	<b>135</b>
<b>4.6.3 Phylogenetic Analysis and Ancestral State Reconstruction .....</b>	<b>135</b>
<b>4.6.4 Constructs for ectopic expression of Alr2 alleles .....</b>	<b>136</b>
<b>4.6.5 Expression of <i>Alr2</i> alleles in mammalian cells.....</b>	<b>137</b>
<b>4.6.6 Aggregation assay .....</b>	<b>137</b>
<b>4.6.7 Sequence Variability and visualization of Domain 1 .....</b>	<b>138</b>
<b>5.0 Extreme variation in gene sequence and copy number in the <i>Alr</i> gene family .....</b>	<b>140</b>
<b>5.1 Foreword .....</b>	<b>140</b>
<b>5.2 Summary .....</b>	<b>140</b>
<b>5.3 Introduction .....</b>	<b>141</b>
<b>5.4 Results.....</b>	<b>142</b>
<b>5.4.1 Assembly of the heterozygous 291-10 ARC reference sequence.....</b>	<b>142</b>
<b>5.4.2 The 291-10 ARC contains a larger set of the Alr family .....</b>	<b>144</b>
<b>5.4.3 <i>Alr</i> content variation between 291-10 and ARC-F .....</b>	<b>153</b>
<b>5.4.4 High allelic polymorphism at <i>Alr1</i>, <i>Alr2</i>, and <i>Alr6</i>.....</b>	<b>161</b>
<b>5.4.5 The region including <i>Alr 3-14</i> was homogenized between haplotypes in inbred lines.....</b>	<b>164</b>
<b>5.5 Discussion .....</b>	<b>165</b>
<b>5.6 Methods .....</b>	<b>168</b>
<b>5.6.1 Sequencing and assembly of the genome of colony 291-10.....</b>	<b>168</b>
<b>5.6.2 RNA extraction and sequencing .....</b>	<b>171</b>

5.6.3 Assembly of the 291-10 ARC and alignment to the ARC-F reference .....	171
5.6.4 Annotation of <i>Alr</i> genes in the 291-10 ARC .....	171
5.6.5 Variation analysis of alleles .....	172
6.0 Discussion and Future Directions.....	173
6.1 The ARC contains a large <i>Alr</i> -like gene family that are novel members of the IgSF .....	173
6.2 <i>Alr2</i> binding specificity can evolve quickly with point mutations in domain 1 ....	175
6.3 <i>Alr1</i> , <i>Alr2</i> , and <i>Alr6</i> are highly polymorphic.....	177
6.4 Future directions .....	179
7.0 Externship research: Cryopreservation of <i>Hydractinia symbiolongicarpus</i> sperm to support community-based repository development for the preservation of genetic resources.....	185
7.1 Foreword .....	185
7.2 Summary .....	185
7.3 Introduction .....	186
7.4 Results.....	188
7.4.1 Sperm motility and viability.....	188
7.4.2 Determining cryoprotectant toxicity to sperm .....	191
7.4.3 Identifying suitable freezing conditions .....	192
7.5 Discussion .....	200
7.5.1 Sperm motility and viability.....	200
7.5.2 Determining cryoprotectant toxicity to sperm .....	202
7.5.3 Identifying suitable freezing conditions .....	203

7.5.4 Approaches to repository development for aquatic species .....	206
7.6 Conclusions .....	208
7.7 Acknowledgements .....	210
7.8 Materials and methods .....	211
7.8.1 Ethics .....	211
7.8.2 Animal care and breeding .....	211
7.8.3 Estimation of sperm concentration and motility .....	214
7.8.4 Longevity and temperature sensitivity of sperm .....	215
7.8.5 Fertility of sperm .....	215
7.8.6 Acute Toxicity of Cryoprotectants .....	216
7.8.7 Standardized Sperm Collection (3-D printing) .....	216
7.8.8 Freezing .....	219
7.8.9 Thawing and use for fertilization .....	220
Appendix A Supplemental Figures .....	221
Appendix B Supplemental Tables .....	225
Bibliography .....	231

## List of Tables

Table 1. Summary of known genes involved in allorecognition. ....	34
Table 2. Overlap coordinates of genomic contigs and BAC contigs used to create reference ARC-F sequence.....	43
Table 3. Gene models classified as <i>Alr</i> pseudogenes. ....	47
Table 4. Sequence homology of Domain 1. ....	64
Table 5. Sequence homology and predicted structural homology of Domain 1.....	66
Table 6. Sequence homology Domain 2 and 3.....	76
Table 7. Sequence homology and predicted structural homology of Domain 2 and Domain 3. .....	78
Table 8. Sequence homology of the ECS fold. ....	86
Table 9. Sequence homology and predicted structural homology of ECS (trimmed).....	88
Table 10. Predicted structural homology of <i>Alr</i> 2 domain 1 isoforms.....	121
Table 11. pI DDT scores of variant positions.....	122
Table 12. Sites in domain 1 under positive or negative selection.....	127
Table 13. <i>Alr</i> annotations in the ARC.....	147
Table 14. <i>Alr</i> bona fide gene classification comparison between the F, pr, and se haplotypes. .....	156
Table 15. <i>Alr</i> putative gene classification comparison between the F, pr, and se haplotypes. .....	157
Table 16. <i>Alr</i> pseudogene classification comparison between the F, pr, and se haplotypes. .....	159

<b>Table 17. Additional <i>Alr</i> annotations found exclusively in the 291-10 haplotype.....</b>	<b>160</b>
<b>Table 18. Amino acid variation in the ectodomain. ....</b>	<b>162</b>
<b>Table 19. Amino acid variation in the cytoplasmic tail. ....</b>	<b>163</b>
<b>Table 20. Overview of frozen samples and fertilization potential. ....</b>	<b>193</b>
<b>Table 21. Slicer software settings used to 3-D print collection chamber. ....</b>	<b>218</b>
<b>Table 22. Printer hardware features. ....</b>	<b>219</b>

## List of Figures

Figure 1. Summary of bacterial allorecognition system in <i>M. xanthus</i> .....	5
Figure 2. Summary of <i>P. mirabilis</i> allorecognition system. ....	7
Figure 3. <i>N. crassa</i> allorecognition system.....	11
Figure 4. Brassicaceae self-incompatibility system.....	16
Figure 5. Solanaceae self-incompatibility system.....	18
Figure 6. Papaveraceae self-incompatibility system. ....	20
Figure 7. Summary of Dictyostelium allorecognition system. ....	23
Figure 8. <i>Hydractinia symbiolongicarpus</i> morphology.....	30
Figure 9. <i>Hydractinia</i> alloresponses.....	32
Figure 10. Alr1 and Alr2 domain architecture. ....	33
Figure 11. Pedigree of colonies used to generate ARC-F reference sequence.....	40
Figure 12. Detail of the initial ARC reference assembly. ....	41
Figure 13. Assembly of the Alr gene complex. ....	42
Figure 14. Annotation of Alr-like genes in the Allorecognition Complex. ....	44
Figure 15. Expression of <i>Alr</i> genes. ....	45
Figure 16. Domain architecture of <i>Alr</i> genes.....	46
Figure 17. Alternative splicing of <i>Alr1</i> and <i>Alr6</i> . ....	49
Figure 18. Alternative splicing at <i>Alr30</i> and <i>Alr2</i> . ....	50
Figure 19. Pairwise amino acid alignment between Alr extracellular domains.....	51
Figure 20. Sequence similarity between Alr extracellular domains.....	52
Figure 21. Neighbor-joining trees of Alr extracellular domains .....	53

Figure 22. Alr1 and Alr2 domain architecture. ....	62
Figure 23. Structural predictions of domain 1. ....	67
Figure 24. Predicted topology of $\beta$ -strands of domain 1. ....	69
Figure 25. Sequence logo of amino acid frequencies in canonical V-set and <i>Alr</i> domain 1. ....	70
Figure 26. CWC “pin” motif comparison.....	71
Figure 27. Salt bridge comparison. ....	72
Figure 28. Hydrophobic core comparison. ....	73
Figure 29. Predicted $\beta$ -turn and hydrogen bonds in Alr D1 structures.....	74
Figure 30. Structural predictions of domains 2 and 3. ....	79
Figure 31. Topology of $\beta$ -strands of I-set and C2 domains. ....	80
Figure 32. Sequence logo of amino acid frequencies in canonical I-set and Alr domain 2 and 3.....	82
Figure 33. Predicted $\beta$ -turn and hydrogen bonds in Alr D2 structures.....	83
Figure 34. Structural predictions of the ECS.....	89
Figure 35. Topology of $\beta$ -strands in Fn3 domains, C1-set Ig domains, and C2-set Ig domains. ....	90
Figure 36. Predicted topology of $\beta$ -strands in the Alr1 and Alr30.3 ECS folds. ....	91
Figure 37. Sequence logo of amino acid frequencies in canonical Fn3 and the Alr ECS.....	92
Figure 38. Invariant cysteines in domains 2, 3, and the ECS fold.....	93
Figure 39. Model of the invariant cysteines in Alr1 domain 2 through the ECS fold. ....	95
Figure 40. Alr cytoplasmic tails are diverse. ....	96
Figure 41. Sequence analysis of cytoplasmic tails and their ITAM and ITIM motifs.....	98

Figure 42. Comparison of human Syk and <i>Drosophila</i> Shark to <i>Hydractinia</i> Syk and Shark. .....	99
Figure 43 Isoform-specific, homophilic binding of Alr2 isoforms.....	109
Figure 44 Relationship of five naturally occurring <i>Alr2</i> alleles.....	110
Figure 45. Plasmid template used in cell aggregation assay and assay results. ....	111
Figure 46. Anc, 046B, and Hap074 versus 214E06. ....	112
Figure 47. Node network of isoforms colored by binding specificity. ....	113
Figure 48. Single-step domain 1 mutants are capable of homophilic binding. ....	114
Figure 49. Anc-T76R and Anc-E93K cell aggregation assay pairwise results. ....	115
Figure 50. Hap074-S44G and Hap074-G47E cell aggregation assay results.....	117
Figure 51. Domain 1 isoforms can evolve via intermediates with broadened specificity. ..	118
Figure 52. Effects of N32Y mutation on binding specificity. ....	119
Figure 53. pLDDT scores of predicted domain 1 structures. ....	123
Figure 54. Predicted structure of Anc domain 1 with the six variant residues labeled in 111A06 and 214E06.....	124
Figure 55. Sequence conservation mapped onto domain 1. ....	125
Figure 56. Residues predicted to have experienced either diversifying or purifying selection mapped onto domain 1. ....	126
Figure 57. Hypothetical binding topologies Alr2. ....	129
Figure 58. Pedigree of the wildtype colony 291-10 used for genomic sequencing.....	142
Figure 59. ARC contig alignment between F reference sequence and the wildtype haplotypes. .....	144
Figure 60. Alr annotations identified in 291-10.....	146



Figure 61. <i>Alr</i> annotations in Clusters. ....	146
Figure 62. Alternative splice variants involving entire exons. ....	149
Figure 63. <i>Alrs</i> with splice junction variants. ....	151
Figure 64. Evidence of exon shuffling in <i>Alr</i> sequences. ....	152
Figure 65. Alignment between ARC-F, ARC-pr, and ARC-se. ....	154
Figure 66. ARC Clusters A, B, C alignment between F, pr, and se haplotypes. ....	155
Figure 67. Variant Alr8 domain architecture. ....	158
Figure 68. Alr19A-se has a duplicate domain 1 in its sequence. ....	160
Figure 69. Pedigree and relationship of the inbred colonies 245-7 (ARC-R homozygous) and 236-21 (ARC-F homozygous). ....	164
Figure 70. The region surrounding Cluster A was homogenized between ARC-R and ARC- F. ....	165
Figure 71. Correlations among sperm velocity, motility, and concentration. ....	189
Figure 72. Estimated sperm fertilization capacity over time. ....	190
Figure 73. Number of fertilized eggs using cryoprotectant-treated sperm. ....	191
Figure 74. Cooling curve for Experiment 1, 5°C/min. ....	194
Figure 75. Cooling curve for Experiment 1, 30°C/min. ....	195
Figure 76. Cooling curve for Experiment 2, 20°C/min. ....	197
Figure 77. Fertilization comparing frozen sperm. ....	198
Figure 78. Cooling curve for variable sperm concentration at 20°C/min. ....	199
Figure 79. Pedigree of the colonies used to generate germplasm and offspring. ....	212
Figure 80. Time lapse of sperm release. ....	213
Figure 81. Sperm collection chamber. ....	217

## Acknowledgements

I would like to express my sincerest gratitude for my mentor Dr. Matthew Nicotra for his invaluable guidance, support, patience, and mentorship throughout the course of my scientific journey and providing me with the opportunity to work in his lab. In addition, I would also like to thank all those who have had an impact on my development as a scientist during my PhD: my committee, Dr. Andrew VanDemark, Dr. Anne-Ruxandra Carvunis, Dr. Jon Boyle, and Dr. Adam Kwiatkowski for their guidance; Dr. Terrence Tiersch and the members of the Aquatic Germplasm and Genetic Resources Center at Louisiana State University for their assistance and friendship during my externship; Dr. Neil Hukriede and the ISB program faculty for their contributions to the program and Shari Murphy for her continual support to the program's students; all members of the Nicotra Lab for their support and friendship; and those in the Immunology, Surgery, and Computational Biology Departments who assisted me with new techniques and provided support to help me achieve my research goals. Lastly, I would like to express my deepest appreciation for the continual support from my family, friends, classmates, and roommate during my time here in Pittsburgh.

## **1.0 Introduction**

### **1.1 Self and non-self recognition in nature**

The ability to distinguish self from non-self tissues is present throughout the tree of life. One type of self/non-self recognition, called allorecognition, is restricted to molecular recognition between individuals of the same species. My dissertation addresses four questions that are relevant to all allorecognition systems and are important to understanding the development and evolution of recognition systems.

First, what genes are involved in allorecognition? Allorecognition systems can function through a single gene or multiple genes. Identifying all genes involved in determining allorecognition specificity (or allotype) is essential to understanding how each contributes to recognition. Comparing the genes responsible for controlling allorecognition to other models throughout the tree of life will help improve our understanding of how these systems developed the various regulatory mechanisms for allorecognition.

Second, what are the structural domains that control allorecognition? The relationship between protein structure and function is key in identifying how allorecognition is controlled at the molecular level. Identifying the structural domains will help determine the molecular interactions that drive allorecognition and provide some insight into any homology between allorecognition systems.

Third, what polymorphisms are present in the allorecognition system? Any system involved in self/non-self recognition must have polymorphism at some level in order to properly distinguish self from non-self. Polymorphism can exist at the nucleotide level, such as with point

mutations, or it can be present at a genomic level, such as with copy number variation. Understanding where polymorphism occurs in a system and how it affects allorecognition will elucidate how new polymorphisms arise.

Fourth, how do new alleles with distinct self-identity specificities evolve? In allorecognition systems, rare alleles with unique specificity are more fit than common alleles in the population. The evolution of new specificities is widely considered to occur through random mutations that result in gain or loss of function. However, most systems studied to address this question have relied on alleles or sequences with numerous differences which make tracing the specific steps which alter specificity difficult to resolve. In addition, many times the only alleles available for analysis are chimeric leaving more uncertainty as to whether the changes in specificity would be biologically relevant. By identifying cases in which stepwise changes can be correlated with the effect on specificity, I can broaden our understanding of the constraints that drive the evolution of those new specificities.

To put into perspective the variety observed in allorecognition systems with respect to the four aforementioned research questions, this introduction reviews the allorecognition systems of several different organisms across the tree of life.

## **1.2 Bacterial allorecognition**

Bacteria use allorecognition to form multicellular groups, in many cases called “swarms”, which can collectively migrate (Stefanic et al., 2015; Cao et al., 2019; Chittor & Gibbs, 2021). Upon forming a swarm, cells cooperate to move across surfaces by synthesizing new flagella and secreting a surface-active agent (“surfactant”) that decreases the surface tension and allows the

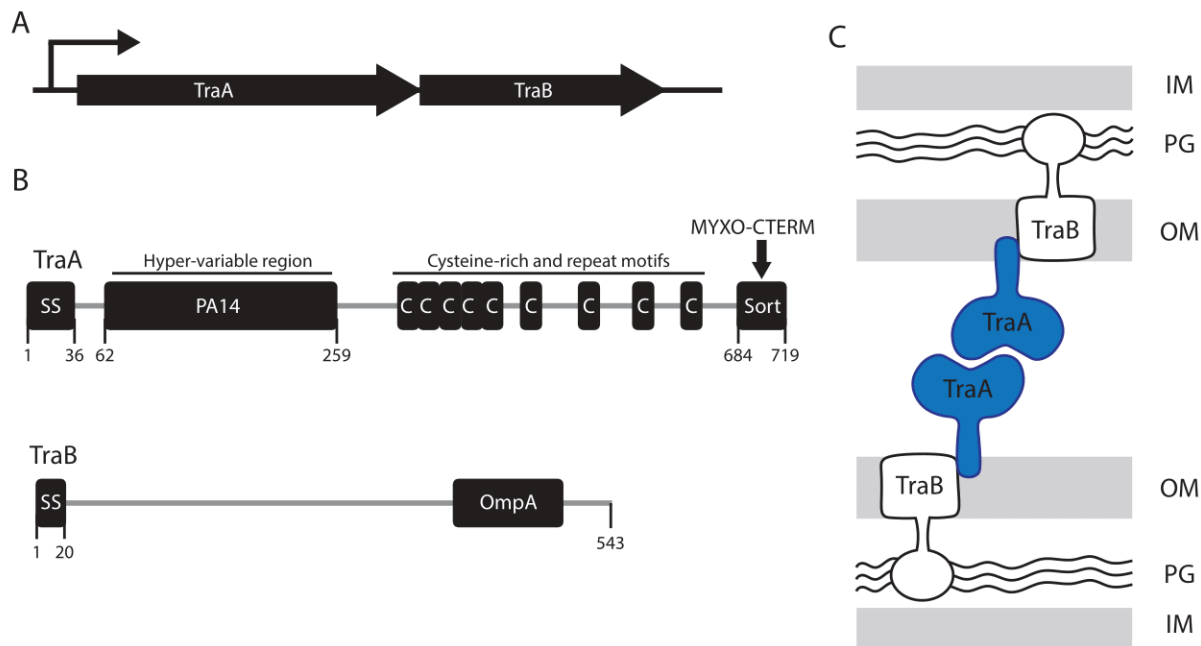
group to migrate quickly (Copeland & Weibel, 2009; Kearns, 2010; Partridge & Harshey, 2013). Single cells usually do not synthesize additional flagella nor produce surfactant because the high metabolic cost to produce these limits them from enhancing their individual motility (Copeland & Weibel, 2009; Kearns, 2010; Partridge & Harshey, 2013). Furthermore, recognition in bacteria is highly specific. As a result, most swarm-forming bacteria are capable of forming strain-specific swarms, also known as the Dienes phenomenon (Dienes, 1946; Senior, 1977; Vos & Velicer, 2009; Stefanic et al., 2015; Lyons et al., 2016; Tipping & Gibbs, 2019).

Cells within a swarm can share resources between kin (Kearns, 2010; Vassallo et al., 2015; Dey et al., 2016). By sharing resources, bacteria can repair their damaged kin by transferring membrane components to repair lethal cell damage and restore cell membrane integrity to benefit the population as a whole (Vassallo et al., 2015). In addition, swarm formation can improve antibiotic tolerance (Benisty et al., 2015; Lyons et al., 2016; Partridge et al., 2018; Troselj et al., 2018). The mechanisms that improve antibiotic tolerance are not fully understood. Some studies have shown that the three dimensional structure of a swarm changes when exposed to antibiotics and horizontal gene transfer (HGT) can impact a swarm's tolerance (Benisty et al., 2015; Partridge et al., 2018).

Even though recognition commonly results in multicellularity, bacterial recognition systems are diverse. Some bacteria determine recognition specificity via a single gene while others require numerous genes. At this point, recognition mechanisms are not easily predicted based on the genomic content of bacteria and must be experimentally validated. Two of the best studied systems are in the swarm-forming bacteria *Myxococcus xanthus* (Cao et al., 2019) and *Proteus mirabilis* (Chittor & Gibbs, 2021).

### 1.2.1 *Myxococcus xanthus*

In *M. xanthus*, a single locus, *traAB*, controls self-recognition (Pathak et al., 2013; Cao & Wall, 2017, 2019; Cao et al., 2019). *traAB* encodes two proteins, TraA and TraB, which are both required for cell-cell adhesion and swarm formation (Cao & Wall, 2017) (Figure 1A). TraA contains a variable domain (VD), cysteine-rich repeats, and a putative MYXO-CTERM motif thought to be involved in protein sorting (Figure 1B). TraB includes an outer membrane  $\beta$ -barrel domain and OmpA cell wall binding domain (Figure 1B). Both TraA and TraB are required for recognition to occur. Only TraA contributes to the recognition specificity between bacteria through homotypic VD interactions between opposing cell membranes (Pathak et al., 2013; Cao & Wall, 2017; Cao et al., 2019) (Figure 1C). TraB likely forms an adhesion complex with TraA through the conserved portion of its sequence in order to anchor TraA in the cell wall (Cao & Wall, 2017, 2019) (Figure 1C). The VD of TraA has distant sequence homology to the PA14 domain (Pathak et al., 2012; Cao et al., 2019). The PA14 domain is found in both eukaryotic and prokaryotic proteins across many species and consists of a  $\beta$ -barrel fold thought to be involved in enzymatic activity or protein binding (Rigden et al., 2004; Cao et al., 2019). No known homologs for *traAB* exist outside of the *Myxococcus* order (Cao et al., 2019).



**Figure 1. Summary of bacterial allorecognition system in *M. xanthus*.**

A) Genomic region encoding TraAB. B) Domain structure of TraA and TraB. SS = Signal Sequence; Sort = sorting tag. C) Putative model for TraA/B interactions to form aggregates amongst bacteria. IM = Inner Membrane; PG = Peptidoglycan; OM = Outer Membrane. Adapted from Pathak et al., 2012 (A, B) and Cao & Wall, 2019 (C)

(<http://creativecommons.org/licenses/by/4.0/>).

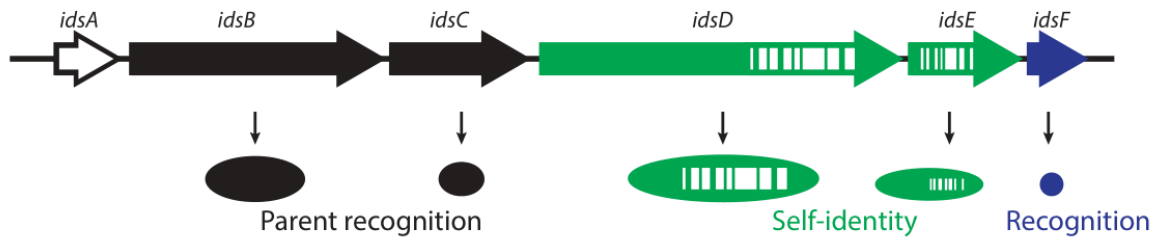
A study surveying 16 TraA alleles from environmental isolates of different *M. xanthus* strains showed that over half of the residues in the VD are polymorphic with many positions encoding more than two amino acid variants (Pathak et al., 2013). These 16 alleles do not each represent a unique recognition specificity. Several of the alleles tested have overlapping specificities yielding only six recognition specificities among the 16 alleles (Pathak et al., 2013; Cao & Wall, 2017; Cao et al., 2019). The evolution of new TraA recognition specificities is hypothesized to occur through chimeric alleles, based on the evidence of HGT in the TraA alleles tested, and amino acid substitutions in the VD (Pathak et al., 2013). The recognition specificity of

TraA can change abruptly with even a single amino acid substitution, also known as a “molecular recognition switch”, as demonstrated in laboratory generated chimeras (Cao & Wall, 2017).

### **1.2.2 *Proteus mirabilis***

In *P. mirabilis*, recognition between swarms is achieved through proteins encoded in the *ids* (identification of self) operon (Gibbs et al., 2008). The *ids* operon encodes six genes (*idsA-F*), of which a secreted protein, *idsD*, and a transmembrane protein, *idsE*, interact to identify closely related kin (Cardarelli et al., 2015; Saak & Gibbs, 2016; Zepeda-Rivera et al., 2018; Tipping & Gibbs, 2019; Chittor & Gibbs, 2021) (Figure 2). The function of *idsA* is unknown but when deleted from the operon does not affect allorecognition responses. *idsB*, *idsC*, and *idsF* are not identity determinants but are required for allorecognition as their products interact with *idsD* and *idsE*, likely as molecular chaperones or in a secretion-dependent manner. The structure, exact localization, and signaling mechanisms of *idsD* and *idsE* remain unknown (Zepeda-Rivera et al., 2018). In order for recognition to occur, *idsD* from one cell must be inserted into a neighboring cell where it interacts with the neighboring cell’s transmembrane protein *idsE* (Saak & Gibbs, 2016; Chittor & Gibbs, 2021). When *idsD* and *idsE* do not bind, swarms will not merge and a boundary remains formed between the swarms (Gibbs et al., 2008, 2011; Wenren et al., 2013; Saak & Gibbs, 2016). Both *idsD* and *idsE* encode their own polymorphic variable region. It is hypothesized that the variable regions of each protein interact and together function as the marker of self (Gibbs et al., 2008) (Figure 2).





**Figure 2. Summary of *P. mirabilis* allorecognition system.**

Genomic architecture of the *ids* operon and putative functions of the various *ids* genes. Variable region of self-identity genes shown as “barcode”. (Adapted from Gibbs et al., 2008. Reprinted with permission from AAAS)

Unlike TraA in *M. xanthus*, the evolution of recognition specificity in *P. mirabilis* requires both *idsD* and *idsE* to co-evolve. If either *idsD* or *idsE* acquire a mutation which prevents them from interacting, a cell will no longer be able to identify self and likely remain separate from any swarms (Pathak et al., 2013; Cardarelli et al., 2015; Hirose et al., 2017). *idsD* also interacts with *idsC*, its molecular chaperone (Zepeda-Rivera et al., 2018), and SdaC, a serine transporter that aids in the insertion of *idsD* into neighboring cells (Chittor & Gibbs, 2021). Although *idsC* and SdaC are not involved in determining the recognition specificity, their interactions with *idsD* places additional constraints on the evolution of *idsD* as mutations that abolish these interactions, whether or not they produce a new recognition specificity, could prevent recognition of kin if *idsD* is not properly trafficked (Chittor & Gibbs, 2021).

### 1.3 Fungal allorecognition

Fungi are a highly diverse and can be classified into three major groups: single celled yeasts, multicellular filamentous molds, and macroscopic filamentous fungi. In molds and fungi, the filamentous growth occurs through thread-like structures called hyphae. Hyphae extend and grow from the apex and can create new tips through a process called branching. When a hyphal filament comes into contact with another filament, they can go through hyphal fusion which is important for increasing growth and developing an interconnected network of hyphae, also referred to as mycelium. Among the allorecognition systems found in eukaryotes, fungi possess two interesting, albeit confusing, allorecognition systems. One system mediates their vegetative compatibility, which regulates the fusion of hyphal cells between fungi (Beadle & Coonradt, 1944; S. J. Saupe, 2000). The second system mediates their sexual compatibility, which necessitates the presence of two opposite mating types in order to reproduce successfully (Coppin et al., 1997; Casselton & Olesnick, 1998; Van der Nest et al., 2014; Kues, 2015). In most species, the allorecognition system that governs vegetative incompatibility does not affect sexual compatibility, and vice versa. In the few rare cases, the allorecognition system controlling sexual compatibility does affect vegetative recognition, but the converse has not been observed (Kwon & Raper, 1967; Newmeyer et al., 1973; Shiu & Glass, 1999).

Sexual compatibility is important in fungi to promote outcrossing and thus limit inbreeding (James, 2015). Fungi are rarely dimorphic, meaning they do not possess distinct sexes based on morphological traits. Instead, fungi possess a “mating type” – the term used to distinguish genetic compatibility between fungi (Blakeslee, 1904). In all cases, successful mating and sexual reproduction only occur between different mating types. Sexual reproduction in fungi can occur through several mechanisms (Kues, 2015). Some species are self-sterile and require two separate

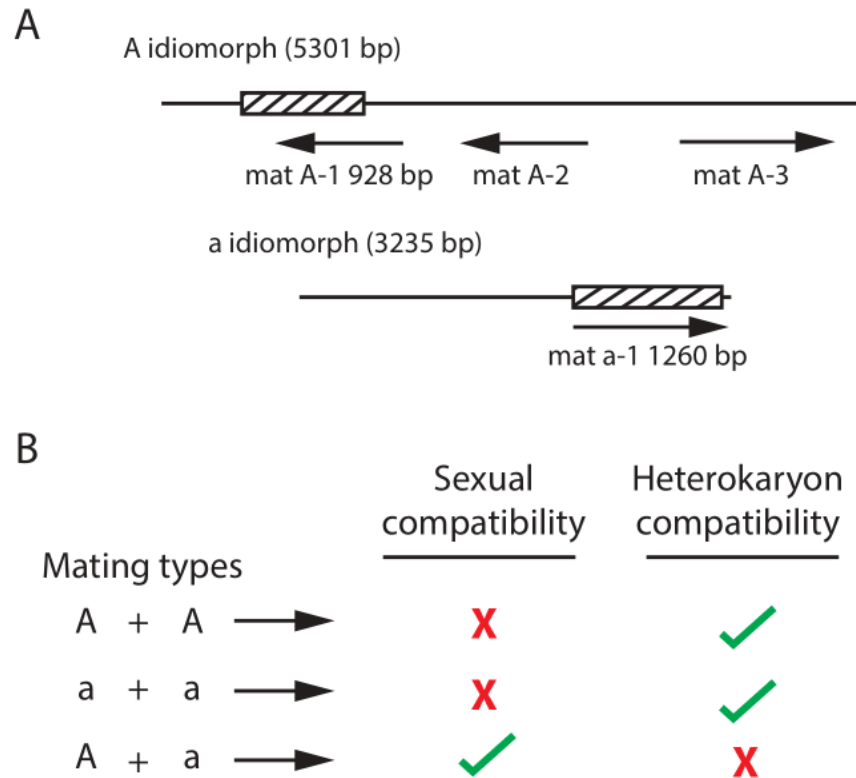
individuals with different mating types in order to reproduce sexually (Coppin et al., 1997). Other species are capable of self-fertilizing as they produce both mating types needed for sexual reproduction (Coppin et al., 1997; Casselton & Olesnick, 1998). The number of mating types in fungi varies drastically. Some species function with only two highly conserved mating types, whereas other species have evolved thousands of mating types (Kues & Casselton, 1992).

Vegetative incompatibility (VI), also referred to as heterokaryon incompatibility (HI), occurs when the hyphal filaments of two fungi grow into each other. The two cells in contact will fuse and become a single cell, commonly referred to as a heterokaryotic cell or heterokaryon (Espagne et al., 2002; Paoletti et al., 2007; Daskalov et al., 2019). If the two genomes of the fused cell are compatible, the heterokaryon will remain fused. Fusion between cells is thought to be important for increasing the size of the colony, improving efficient use of nutrients, and thus increasing the reproductive potential (Bastiaans et al., 2015). If they are incompatible, the heterokaryon will be cut off or compartmentalized and then rapidly killed by a programmed cell death mechanism (S. J. Saupe, 2000; Glass & Kaneko, 2003; Daskalov et al., 2019). The genes that control HI/VI are named *het/vic* (Glass et al., 1988; Leslie, 1993); fungi have numerous *het/vic* loci, usually between 5 and 12 (Bernet, 1967; Cortesi & Milgroom, 1998; Van der Nest et al., 2014). Compatible heterokaryons require all *het* loci to be identical. It has been shown that even a single amino acid difference in one locus results in an incompatible heterokaryon (S. Saupe et al., 1995; Paoletti et al., 2007). HI/VI is thought to be important in preventing resource plundering (Buss, 1982; Debets & Griffiths, 1998) and the transmission of harmful cytoplasmic elements between fused fungi (Bastiaans, Debets, & Aanen, 2015; Bastiaans et al., 2014; Caten, 1972). Additionally, HI/VI may induce the sexual cycle in fungi as a result of encountering non-self during vegetative growth (Dyer et al., 1992).

The presence of these two distinct allorecognition systems in fungi, combined with the genetic variation between species, yields numerous possible distinct recognition systems. Two models that represent some of that diversity in fungal systems are *Neurospora crassa*, an ascomycete and type of red bread mold, and *Schizophyllum commune*, a basidiomycete also known as the split gill mushroom.

### **1.3.1 *Neurospora crassa***

In *Neurospora crassa*, sexual compatibility is determined by a single genetic locus, *mat*, with one of two highly conserved mating types, *A* and *a*, which are both required to achieve successful sexual reproduction (Glass et al., 1990; Staben & Yanofsky, 1990; Chang & Staben, 1994). Due to the mating types *A* and *a* being dissimilar in sequence and evolutionarily unrelated, they are not referred to as alleles but are instead called idiomorphs (Metzenberg & Glass, 1990; Dyer et al., 1992). The *A* idiomorph is comprised of three open reading frames (ORFs), *mat A-1*, *mat A-2*, and *mat A-3* (Figure 3A). Of the three, only *mat A-1* is required for determining mating identity while the other two are only necessary for post-fertilization functions (Glass & Lee, 1992; Ferreira et al., 1998; Shiu & Glass, 1999). The *a* idiomorph only encodes one ORF, *mat a-1*, which is required for both mating identity and post-fertilization functions (Staben & Yanofsky, 1990) (Figure 3A). Both *mat A-1/a-1* genes appear to encode a transcriptional factor that functions in DNA-binding and are suspected to control pheromone and pheromone receptor expression in order to attract the opposite mating type for sexual reproduction (Kues & Casselton, 1992; Philley & Staben, 1994). In this case, it is proposed that the *a* mating type would regulate the expression of the *a* pheromone and *A* pheromone receptor and vice versa in the *A* mating type in order to attract each other (Glass et al., 1990; Philley & Staben, 1994).



**Figure 3. *N. crassa* allorecognition system.**

A) A and a idiomorph from *N. crassa* both required for sexual compatibility (Adapted from Glass et al., 1990). B)

Mating type combinations needed for sexual and heterokaryotic incompatibility responses. Heterokaryon incompatibility is also dependent on the het loci.

HI in *N. crassa* is controlled by at least 11 unlinked *het* loci that must match in order for two fungi to successfully fuse (Mylyk, 1975; Perkins, 1988; Muirhead et al., 2002). The *het* loci that have been molecularly identified thus far are distinct in sequence and are polymorphic with two or three different alleles at each gene (Jennifer Wu et al., 1998; Hall et al., 2010; Zhao et al., 2015). In addition, *N. crassa* is one of the rare fungi that exhibits mating type associated vegetative incompatibility (Newmeyer et al., 1973; Glass et al., 1990; Staben & Yanofsky, 1990; Shiu &

Glass, 1999) (Figure 3B). When hyphae from opposite mating types fuse to form a heterokaryon during vegetative growth, even if the *het* loci are identical, the heterokaryon's growth is inhibited and cell death occurs (Vellani et al., 1994; Shiu & Glass, 1999). This limits vegetative fusion to fungi with the same mating type and the same *het* loci. Many of the *het* genes encode proteins possessing a domain of unknown function, termed HET (Glass & Dementhon, 2006; Zhao et al., 2015; Daskalov et al., 2017). This domain does not appear to be isolated to the known *het* loci but appears at least 69 other times in the genome (Zhao et al., 2015). The HET domain may have similarities to known domains involved in other immune systems, such as the Toll/interleukin-1 receptor (TIR) and the nucleotide-binding oligomerization (NOD)-like receptor (NLR) (Dyrka et al., 2014; Gonçalves & Glass, 2020). Two identified *het* loci, *het-6* and *het-c*, appear on the same chromosome approximately 150 kb apart. While allelic interactions appear to play a primary role in HI, evidence that non-allelic interactions (between different genes) affect HI has also been reported (Kaneko et al., 2006). The molecular mechanisms for allelic and non-allelic recognition have not yet been elucidated. The evolution of new alleles and the maintenance of diversity in this HI system may function to counteract the selection against cheaters, a cell harboring a variant that contributes less to colony functions and more to reproduction (Czárán et al., 2014; Bastiaans et al., 2015).

### **1.3.2 *Schizophyllum commune***

In *Schizophyllum commune*, mating type is controlled by two unlinked loci, *mat A* and *mat B*, which are both highly polymorphic with approximately 288 alleles at *mat A* and 81 alleles at *mat B* (Raper et al., 1958; Kues, 2015). The number of distinct mating types in *S. commune* is estimated to be more than 18,000. The evolution of so many distinct mating types is thought to

promote outbreeding (James, 2015). The *mat A* locus is spread out between two subloci, *Aα* and *Aβ*, which are separated by approximately 550 kb. In total, the *mat A* locus encodes numerous homeodomain proteins that likely function in DNA binding to regulate expression. The *mat B* locus is also split into two linked loci, *Bα* and *Bβ* (Specht, 1995), which both encode numerous proteins associated with DNA expression regulation and pheromone receptor genes. It remains unclear thus far where the *mat A* and *mat B* loci are located in relation to each other (Ohm et al., 2010). Similar to sexual attraction in *N. crassa*, pheromones and pheromone receptors in *S. commune* may also mediate the actual mating type recognition between fungi by expressing a pheromone of the same mating type and a receptor of a different mating type. It is unknown whether all the pheromones and receptors encoded must all be expressed in order for mating type recognition to occur or whether it depends only on some of them.

HI in *S. commune*, as well as most other basidiomycetes, relies on fewer *het* loci than in most ascomycetes. There are at least 5 *het* homologs, which are all multi-allelic (Van der Nest et al., 2014). The location and organization of these in the genome is unknown. The evolution and maintenance of new alleles in these genes is unknown. Based on the polymorphism of these genes, a multitude of allelic combinations likely exist, which would make fungal recognition very specific to themselves and other very closely related fungi. Similar to other fungi, some of the *het* homologs possess conserved domains that have been implicated in HI, such as HET, WD40, and NACHT, which have all been observed in other genome locations as well. The WD40 domain has been implicated in various functions such as signal transduction, transcription regulation, and apoptosis – all of which are necessary processes for HI. The NACHT domain is found in enzymes related to apoptosis. In contrast to *N. crassa*, HI in *S. commune* is not affected by the mating type locus.

## 1.4 Plant allorecognition

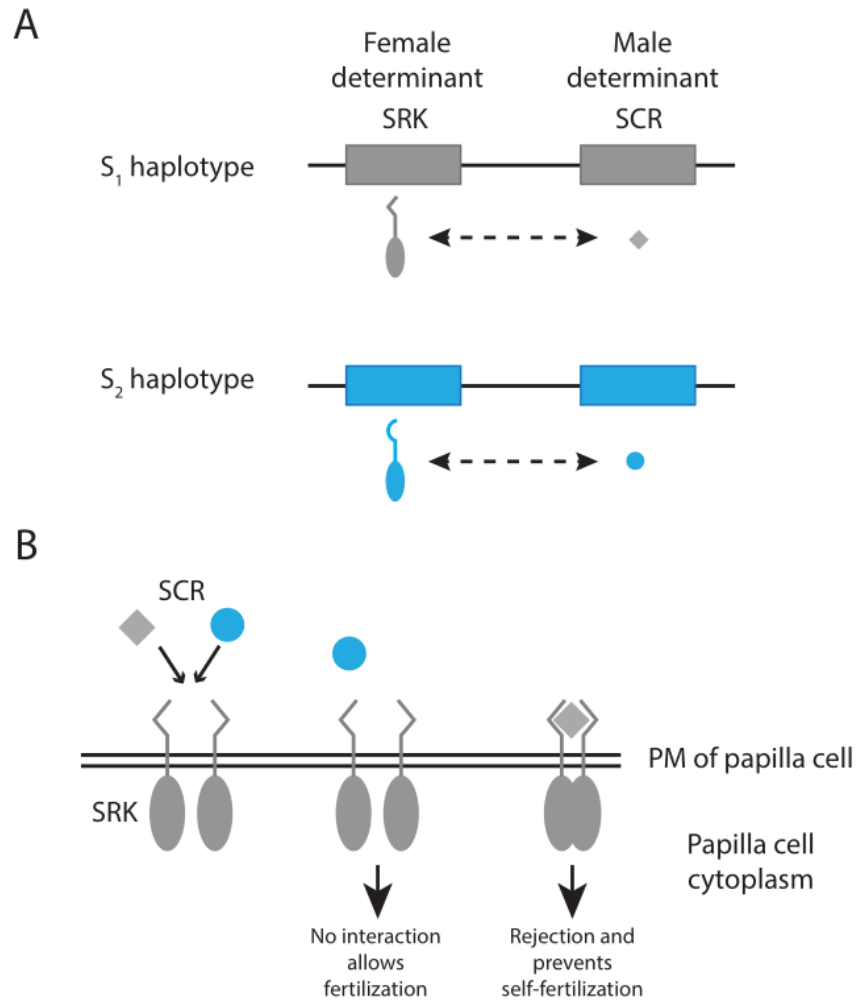
Allorecognition in plants is termed self-incompatibility (SI) and is vital to preventing self-fertilization (Wright, 1939; Bod'ova et al., 2018). Self-fertilization always leads to populations with lower genetic diversity, which could make future propagation difficult in the variable environments that plants face in nature (Rea & Nasrallah, 2008). This is especially apparent when a deleterious allele arises that reduces the fitness of an individual. In this case, the reduced fitness of inbred individuals is referred to as inbreeding depression. Thus, SI in plants promotes outcrossing in order to avoid the negative consequences of inbreeding.

SI is common to many flowering plants (Igic et al., 2008) and usually functions through male-specificity and female-specificity determinants, called S-determinants, which are both encoded at a single polymorphic “S locus” in most species (Takayama & Isogai, 2005; Iwano & Takayama, 2012). The “S locus” is a generic term used to describe this specificity region among the plant SI systems and is not intended to imply homology between S loci. There are three molecular mechanisms that have been described in plants based on the different S-determinants encoded at the S locus (Iwano & Takayama, 2012; Fujii et al., 2016). The three models that have been studied in the greatest depth and best illustrate these molecular mechanisms are: Brassicaceae, the mustard plant family (Takayama & Isogai, 2005); Solanaceae, the nightshade or potato family (Entani et al., 2003; Ushijima et al., 2003; McClure, 2009; X. Meng et al., 2011); and Papaveraceae, the poppy family (M. J. Wheeler et al., 2009, 2010).



### 1.4.1 Brassicaceae

Brassicaceae is one of the more commonly studied families, with over half of the species exhibiting SI (Nasrallah, 2019). In Brassicaceae, the S-locus contains two genes involved in determining specificity, named *S-locus cysteine-rich* (*SCR*) (Stein et al., 1991; Takayama & Isogai, 2005) and *S-locus receptor kinase* (*SRK*) (Schopfer et al., 1999; Suzuki et al., 1999; Takayama et al., 2000; Takayama & Isogai, 2005) (Figure 4A). *SCR* encodes the male determinant for SI which is a secreted cysteine-rich protein that gathers on the pollen surface (Takayama & Isogai, 2005). *SRK* encodes the female determinant for SI, which is a transmembrane protein that contains an extracellular S-domain responsible for ligand binding (Kemp & Doughty, 2007; Naithani et al., 2007; Xing et al., 2013) and an intracellular serine/threonine kinase domain, thought to be involved in signaling (Stein et al., 1991; Stone et al., 1999, 2003). The SRK protein is expressed on the surface of the stigma of the plant and interacts with the SCR protein present in the pollen in a haplotype-specific manner. If SRK and SCR bind in a self-recognition interaction, autophosphorylation of SRK occurs and triggers a signaling cascade that results in rejecting the self-pollen and preventing self-fertilization (Takayama & Isogai, 2005; Nasrallah, 2019) (Figure 4B). The exact residues that determine the binding specificity between SRK and SCR are not yet known (Chookajorn et al., 2004; Nasrallah, 2019). SCR and SRK sequences are both highly variable between alleles as is expected of proteins controlling recognition specificity (Stein et al., 1991; Takayama et al., 2000; Watanabe et al., 2000; Shiba et al., 2002; Takayama & Isogai, 2005).



**Figure 4. Brassicaceae self-incompatibility system.**

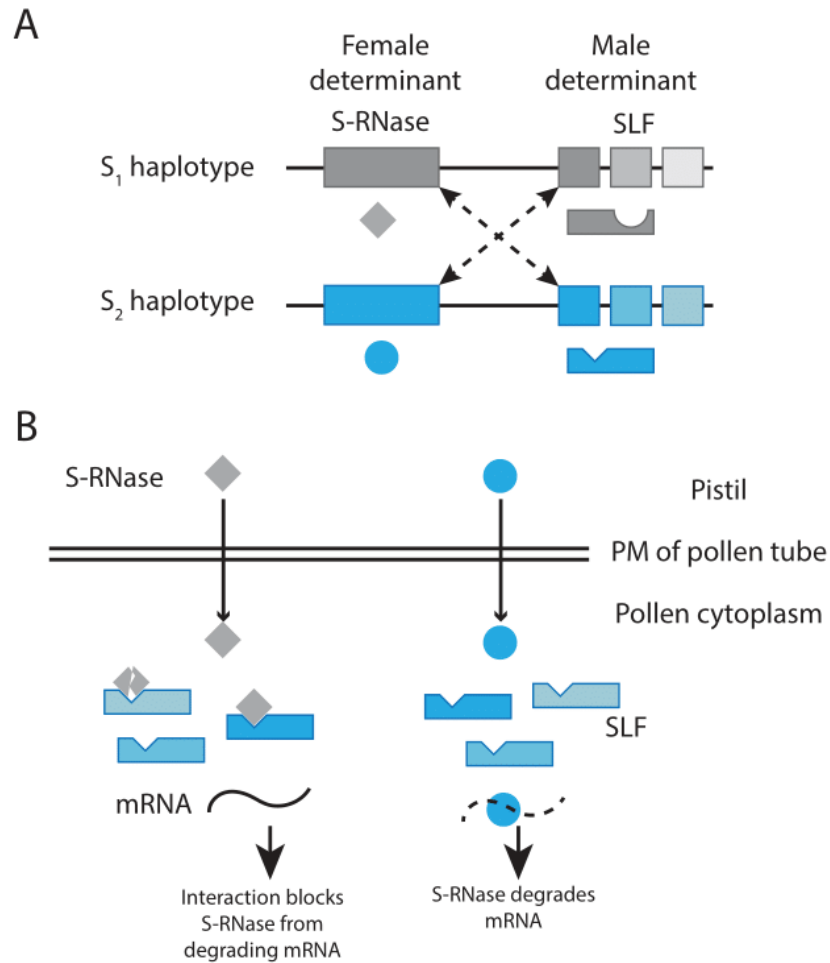
A) Representation of SRK and SCR encoded in the genome from two S haplotypes. B) Model illustrating how SCR and SRK binding prevents fertilization while lack of binding is permissive for fertilization. PM = plasma membrane.

(Adapted from Iwano & Takayama, 2012.)

### 1.4.2 Solanaceae

In Solanaceae, the S locus encodes an *S-RNase* gene and 16 to 20 *SLF* genes (S-locus F-box) (Takayama & Isogai, 2005) (Figure 5A). S-RNase, the female determinant (Lee et al., 1994; Murfett et al., 1994), degrades pollen RNA (Ioerger et al., 1990; Matton et al., 1999; Newbigin et

al., 2008) (Figure 5B). The SLFs are considered the male determinant and encode the F-box family of proteins, which are thought to act as an E3 ubiquitin ligase subunit which mediates ubiquitination and degradation of non-self-S-RNases (D. Wheeler & Newbigin, 2007; Ken-ichi Kubo et al., 2010; Williams et al., 2014, 2015; Ken-inchi Kubo et al., 2015; P. Sun et al., 2015; Li et al., 2017; L. Wu et al., 2018; Vieira et al., 2019). When the style is pollinated, S-RNase enters the pollen tubes and degrades its own pollen RNA, acting as a highly selective cytotoxin (Hua et al., 2008; Muñoz-sanz et al., 2020) (Figure 5B). To counteract this SI mechanism and to allow for the most outcrossing events, the SLFs must identify and ubiquitinate as many of the possible non-self S-RNases so that fertilization can occur with non-self pollen. SLFs do not recognize or degrade self-S-RNase. In the event of self-pollination, because the SLFs do not degrade self-S-RNase, the S-RNase can degrade mRNA and thereby eliminate any self-pollinated cells from growing (Figure 5B). The *S-RNase* gene is highly polymorphic whereas the *SLFs* are less polymorphic and have fewer alleles than *S-RNase*, which accounts for their function in degrading non-self *S-RNase* (Uyenoyama et al., 2001; Newbigin et al., 2008; McClure, 2009). Structural and phylogenetic analyses have identified at least 16 amino acid residues in the various SLF isoforms that may determine the interaction specificity for non-self S-RNases (Vieira et al., 2019).



**Figure 5. Solanaceae self-incompatibility system.**

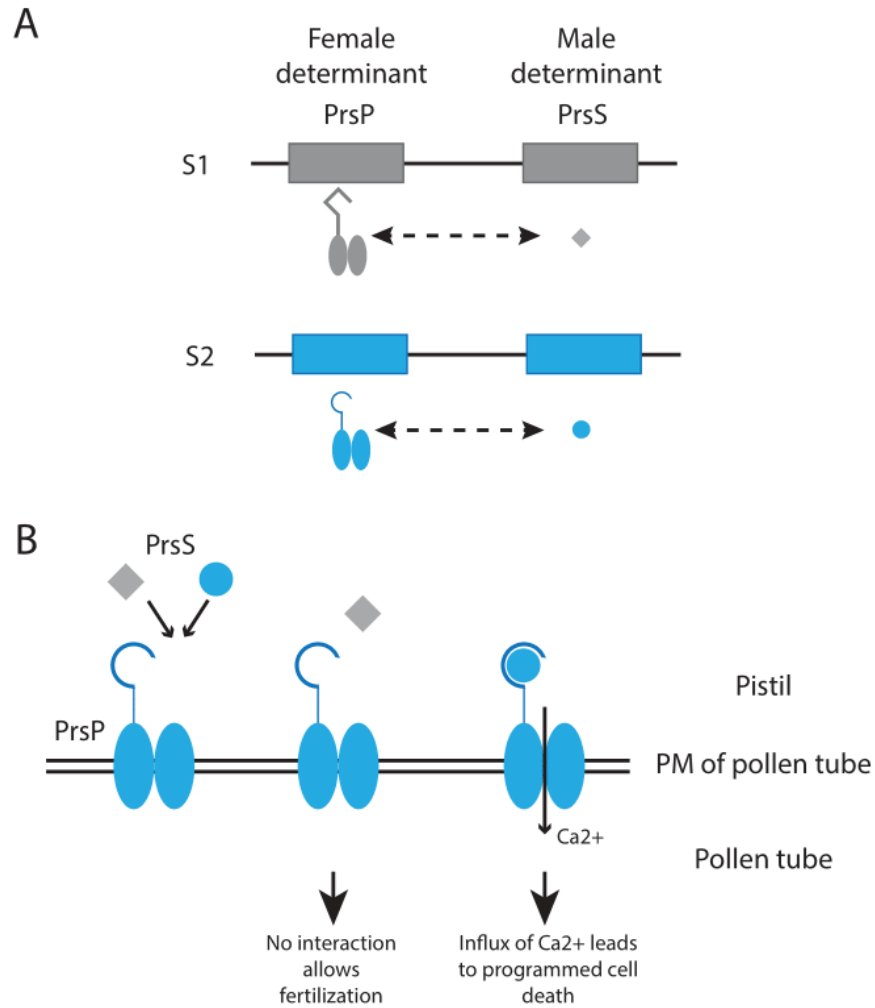
A) Representation of the S-RNase and SLFs encoded in the genome from two S haplotypes. B) Model illustrating how S-RNase and SLF interact to allow fertilization between different S haplotypes. PM = plasma membrane.

(Adapted from Iwano & Takayama, 2012)

### 1.4.3 Papaveraceae

In the Papaveraceae SI system, the female S-determinant, *P. rhoeas* style *S* (*PrsS*), is a small protein that is secreted by the stigmatic papilla cells and acts as a signaling ligand (M. J. Wheeler et al., 2010; L. Wang et al., 2019) (Figure 6A). The male S-determinant, *P. rhoeas* pollen

*S* (*PrpS*), is a transmembrane protein that localizes in the pollen tube (M. J. Wheeler et al., 2009; Iwano & Takayama, 2012) (Figure 6A). In the case of self-pollination, PrsS interacts with an extracellular loop from PrpS inside the pollen tube which initiates SI (Figure 6B). There are multiple downstream effects from this self-interaction between PrsS and PrpS, such as increase in intracellular  $\text{Ca}^{2+}$  and depolymerization of the actin cytoskeleton, which ultimately lead to programmed cell death (Thomas & Franklin-Tong, 2004; M. J. Wheeler et al., 2010; Juyou Wu et al., 2011; L. Wang et al., 2019). PrsS is highly polymorphic with roughly 66 haplotypes estimated (Lane 1993) and PrpS is likewise polymorphic with three alleles sequenced being 40-50% divergent (M. J. Wheeler et al., 2009).



**Figure 6. Papaveraceae self-incompatibility system.**

A) Representation of PrsP and PrsS encoded in the genome from two S haplotypes. B) Model illustrating how PrsS and PrsP interact to prevent self-fertilization through programmed cell death. PM = plasma membrane. (Adapted from Iwasno & Takayama, 2012)

#### 1.4.4 Evolution and selection of S-determinant alleles in plant models

In all models, the SI system is controlled by tightly linked genes defined as the male- and female-determinants (Nasrallah, 2019). All systems rely on polymorphism to prevent self-fertilization while promoting outcrossing (Takayama & Isogai, 2005). The mechanism that drives

SI in Brassicaceae and Papaveraceae relies on a self-recognition interaction between the male and female determinants in order to prevent self-fertilization. The evolution of new S-haplotypes is fairly straightforward in Brassicaceae and Papaveraceae as both determinants must co-evolve in order to maintain their self-interaction (Watanabe et al., 2000; Shiba et al., 2002; Takayama & Isogai, 2005; Iwano & Takayama, 2012). Unlike Brassicaceae and Papaveraceae, Solanaceae SI is based on the non-self interaction between an SLF and a foreign S-RNase so that it may be tagged and degraded before it can degrade the foreign pollen RNA necessary for outcrossing.

Nearly all systems that maintain high levels of diversity are driven by balancing selection, specifically negative frequency-dependent selection (NFDS) (A. D. Richman & Kohn, 2000; A. Richman, 2000; Lawrence, 2000; Charlesworth, 2006; Bod'ova et al., 2018). By maintaining numerous alleles at low or extremely rare frequencies in the population, they have a selective advantage over alleles at higher frequencies (Wright, 1939). The evolution of new specificities in Solanaceae does not appear to be under co-evolution and is still somewhat of an evolutionary puzzle (Newbigin et al., 2008; Ken-ichi Kubo et al., 2010). S-RNases have been shown to have dual specificities, which allow them to reject self and closely-related non-self S-RNases and to maintain their old recognition phenotype while evolving a new unique specificity (Matton et al., 1999). Another possibility is that an S-RNase evolves to have a weaker affinity with its own SLFs so that it can avoid degradation and thereby be able to degrade pollen RNA (X. Meng et al., 2011).

### **1.5 Slime mold allorecognition**

In *Dictyostelium discoideum*, a social amoeba also referred to as a slime mold, allorecognition plays a key role in the organism's development and sociality. When food sources

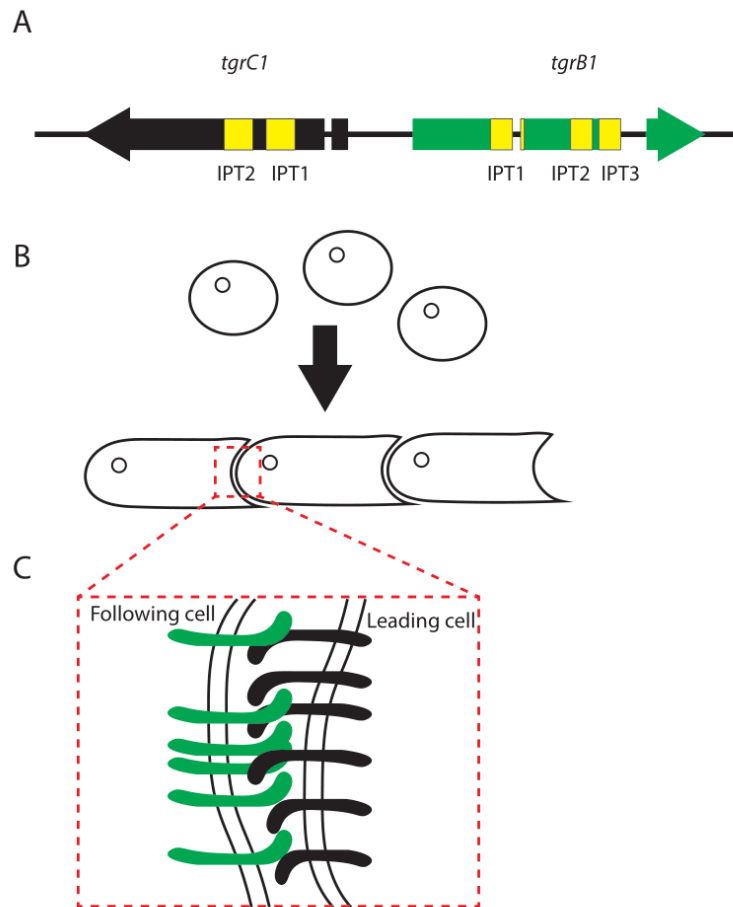
are plentiful, *Dictyostelium* remain unicellular and can divide through fission, whereas when food becomes scarce, *Dictyostelium* cells will form aggregates and then go through their developmental cycle (Kolbinger et al., 2005). These aggregates will behave as multicellular organisms and are capable of migrating based on optimal environmental conditions.

During their developmental cycle, the multicellular *Dictyostelium* differentiate its cells into the various cell types that make up the fruiting body structure. The fruiting body structure consists of the stalk, which will ultimately die, and the spores, which are the viable offspring of the fruiting body and are the only cells that pass their genes on to the next generation. Because of this, cheaters present within the population can preferentially differentiate into spore cells, thereby accessing the benefits without contributing to the costs involved in aggregate formation, while leaving the non-cheater cells within the chimera to die in the stalk (H.-I. Ho et al., 2013). Allorecognition prevents cheaters from invading by only allowing kin to aggregate and form the fruiting body.

Allorecognition is controlled by two genes, TigerC1 (*tgrC1*) and TigerB1 (*tgrB1*), which are encoded directly next to each other (Dynes et al., 1994; J. Wang et al., 2000; Benabentos et al., 2009; Hirose et al., 2011; Kundert & Shaulsky, 2019) (Figure 7A). Both genes have similar structural predictions with a single-pass transmembrane domain, multiple extracellular immunoglobulin-like domains, and cytoplasmic tails. However, *tgrC1* and *tgrB1* differ at the protein level with ~13% differences between the extracellular domains. The *tgrB1* and *tgrC1* proteins heterophilically bind to each other in an isoform-specific manner between cells (Chen et al., 2013, 2014; Hirose et al., 2017; Kundert & Shaulsky, 2019). Binding occurs between *tgrC1* from the tail of the leading cell to *tgrB1* from the head of the following cell (Fujimori et al., 2019) (Figure 7B,C). Although *tgrC1* and *tgrB1* form a heterophilic interaction, a study tested the interactions between five different *tgrC1* and *tgrB1* allotype pairs and found that only *tgrC1* and



*tgrB1* from the same allotype were capable of binding (Hirose et al., 2011). When populations of cells with different *tgr* allotypes are mixed, they begin to segregate around 8 hours of development and develop into fruiting bodies (Hirose et al., 2011). The aggregates, and subsequent fruiting bodies, do not always completely separate as few incompatible cells may be scattered within a largely homogenous aggregate (Ostrowski et al., 2008; H. Ho & Shaulsky, 2015).



**Figure 7. Summary of Dictyostelium allorecognition system.**

A) *tgrC1* and *tgrB1* are encoded directly next to each other in a head-to-head orientation. Yellow indicates distinct immunoglobulin domains. B) Individual cells interact and orient into a head-to-tail orientation. C) Model for interaction between *tgrC1* and *tgrB1* between cells. (Adapted from Kundert & Shaulsky, 2019)

Both *tgrC1* and *tgrB1* are highly polymorphic. One genetic study reported that in 30 alleles of *tgrC1*, 319 codons (out of ~880) were found to be polymorphic with up to 7 variants per codon; in 29 alleles of *tgrB1*, 266 codons (out of ~900) were polymorphic with up to 8 variants per codon (Benabentos et al., 2009). *tgrC1* and *tgrB1* appear to be under balancing or positive selection. For instance, *Dictyostelium* will not develop properly with incompatible *tgrB1* and *tgrC1*, suggesting that *tgrB1* and *tgrC1* may co-evolve (Kundert & Shaulsky, 2019).

How new alleles evolve within this system has not been determined yet. One hypothesis is that if a duplication event results in two copies of *tgrB1* and *tgrC1*, the multiple alleles will allow new alleles to evolve while retaining the parental allotype. An alternative hypothesis incorporates the observation that a gain of function allele in *tgrB1* which eliminates the need for *tgrC1* for *Dictyostelium* to develop successfully and could relax the selective pressure enough to allow for co-evolution of *tgrC1* (Hirose et al., 2017).

## **1.6 Invertebrate allorecognition**

Allorecognition is common to numerous colonial marine invertebrates and prevents fusion between conspecifics. Although it is highly prevalent in these animals, very few genetic systems have been characterized. One prevailing hypothesis as to why allorecognition is present in so many animals is that it is important for preventing germ cell parasitism (Buss, 1982, 1987; Rinkevich et al., 1992; Stoner & Weissman, 1996; Stoner et al., 1999; Laird et al., 2005; De Tomaso, 2006). In most models, allorecognition allows the two animals to fuse and form a common vasculature through which they can share resources and cells throughout the expanded colony. In nature, fusion is a relatively rare event and a rejection response is more common based on highly restricted kin

selection (Nicotra & Buss, 2005). Three models for invertebrate allorecognition are the tunicate *Botryllus schlosseri*, the sponge *Amphimedon queenslandica*, and the cnidarian *Hydractinia symbiolongicarpus*.

### 1.6.1 *Botryllus schlosseri*

*Botryllus*, also known as the star tunicate, is composed of an outer tunic encompassing individual zooids arranged in a star-like pattern (Milkman, 1967; Rosengarten & Nicotra, 2011). The zooids within the same tunic boundary are termed a system, and a colony of systems uses allorecognition to fuse and thereby expand. A colony can expand and grow to contain thousands of systems that are all connected by a common circulatory system. The circulatory system ends at the colony borders in bulbous extensions called ampullae. When two *Botryllus* colonies are near each other and their ampullae come into contact the colonies will exhibit either a fusion alloresponse, in which the circulatory system will fuse between the colonies, or a rejection response, in which the interacting ampullae are destroyed (Oka & Watanabe, 1960; Mukai & Watanabe, 1975; McKittrick & De Tomaso, 2010; Taketa & De Tomaso, 2015).

In *Botryllus*, allorecognition is controlled through the Fusion-HistoCompatib*ility* (*FuHC*) locus, which encodes at least five genes involved in allorecognition. The first candidate allorecognition gene, *cFuHC* (De Tomaso et al., 1998, 2005; De Tomaso & Weissman, 2003), was ultimately found to encode two genes, *sFuHC* and *mFuHC*, which are located right next to each other (<250 bp apart). *sFuHC* encodes a secreted protein and *mFuHC* encodes a membrane-bound protein (De Tomaso et al., 2005; Nydam et al., 2013). Both genes appear to contribute to the alloresponse; however, they do not always accurately predict the histocompatibility outcome based on their haplotype (De Tomaso et al., 2005; Voskoboynik et al., 2013). The next two genes

identified to be involved in the alloresponse, *fester* and *uncle fester*, are located 150-300 kb upstream of *sFuHC/mFuHC* (Nyholm et al., 2006; McKittrick et al., 2011). Both genes encode putative transmembrane bound receptors. *fester* appears to be important for initiating the rejection response and may be the receptor for the *cFuHC* ligand (Nyholm et al., 2006; Taketa & De Tomaso, 2015). Likewise, *uncle fester* is involved in initiating the rejection response (McKittrick et al., 2011). The fifth gene identified was termed the *Botryllus histocompatibility factor* (*BHF*) and is approximately 62 kb upstream of *sFuHC/mFuHC* and 95 kb downstream of *uncle fester* (Voskoboynik et al., 2013; Taketa & De Tomaso, 2015). *BHF* is predicted to encode a cytoplasmic protein, although no domains have been predicted yet (Letunic et al., 2012; Voskoboynik et al., 2013). The *BHF* haplotype does appear to be highly correlated to the histocompatibility outcome, although it does not predict it perfectly (Voskoboynik et al., 2013). *BHF* is composed of three exons and has two alternatively spliced transcripts. The first isoform is composed of all three exons resulting in a 252 amino acid protein (Taketa & De Tomaso, 2015). The second isoform is composed of exon 1 and an extended exon 2, which results in a 219 amino acid protein (Taketa & De Tomaso, 2015). *BHF* may represent a general tunicate allorecognition factor (Werner A Mueller & Rinkevich, 2020). Because of the lack of predicted domains in *BHF*, particularly those expected for a cell-surface recognition protein, it is possible that further study may elucidate a new mechanism for allorecognition control (Voskoboynik et al., 2013; Taketa & De Tomaso, 2015).

Three of the genes (*mFuHC*, *sFuHC*, and *BHF*) are highly polymorphic with potentially hundreds of alleles in nature (Mukai & Watanabe, 1975; V L Scofield et al., 1982; Virginia L Scofield et al., 1982; Grosberg & Quinn, 1986; Rinkevich et al., 1992; Yund & Feldgarden, 1992; De Tomaso et al., 2005; Nydam & De Tomaso, 2012; Voskoboynik et al., 2013). The polymorphism appears to be maintained by balancing selection in the case of *mFuHC* and *sFuHC*

(Nydam et al., 2017). *fester* and *uncle fester* are more closely related to each other, with several of their exons suggesting a duplication event, and are polymorphic but less so than the *cFuHC* genes (Taketa & De Tomaso, 2015). However, both *fester* and *uncle fester* have numerous alternatively spliced isoforms (between 8-24 in *fester*) in each individual. The polymorphism in *fester* appears to be controlled by either balancing or purifying selection (Nydam & De Tomaso, 2012).

### **1.6.2 *Amphimedon queenslandica***

Sponges are the most primitive multicellular animals alive today. Allorecognition in sponges is not clearly understood on the genetic and molecular level. However, numerous proteins with homology to those in vertebrates appear to be involved with cell-adhesion and signaling functions that would be expected for recognition molecules (Xavier Fernandez-Busquets & Burger, 1999). One of the first studies detailing allorecognition in sponges focused on studying cell adhesion of mechanically dissociated cells (Wilson, 1907). Dissociated cells are capable of movement through pseudopodia until they come into contact and adhere to one another. As more cells moved into contact, the aggregate would continue to grow and reorganize itself until a new miniature sponge was formed (Gaino et al., 1999). In this study, only cells from the same species were capable of forming aggregates (Wilson, 1907). Some studies using other sponge species have observed dissociated cells that begin to form some mixed aggregates before completely or partially sorting out into more monospecific aggregates (Humphreys, 1970a, 1970b; McClay, 1971). However, these observations are thought to be due to either nonspecific adhesive forces or a temporary suppression of species-specific recognition after undergoing dissociation (McClay, 1971). Species-specific cell recognition in sponges is mediated by very large proteoglycan macromolecules termed aggregation factors (AF), which can be extracted from the cell surface of

dissociated cells (Humphreys, 1963; Leith, 1979). AFs can exist in linear form or in a sunburst shape that form complexes that interact between sponges (Dammer et al., 1995; X Fernandez-Busquets & Burger, 2003). Cell adhesion in many AFs occurs through a two-step process whereby each AF forms  $\text{Ca}^{2+}$ -dependent homophilic interactions with itself and  $\text{Ca}^{2+}$ -independent heterophilic interactions with the appropriate cell surface receptors (Jumblatt et al., 1980; Xavier Fernandez-Busquets & Burger, 1999; Grice et al., 2017).

In *A. queenslandica*, there are five clustered genes, *AqAFA-AqAFE*, which are tightly packed within an 80 kb region and possess sequence similarities to AF and AF-like sequences found in other sponges (Grice et al., 2017). All *AqAF* genes are predicted to encode secreted proteins except for *AqAFE*, which lacks a predicted signal peptide but would otherwise be extracellular (Grice et al., 2017). Each *AqAF* has multiple domains that vary in number and sequence between the AFs. All *AqAFs* encode between 2 and 14 Calx- $\beta$  domains, which function in calcium binding, that are highly variable between AFs (average of 25% identity) (Grice et al., 2017). *AqAFB* and *AqAFE* both encode a single von Willebrand type A (VMA) domain, which are very dissimilar (average 31% identity) and *AqAFC* and *AqAFD* each encode a single von Willebrand type D (VMD) domain, which are also very dissimilar (average 23% identity) (Grice et al., 2017). The VMA and VMD domains are found in proteins with a variety of functions including cell adhesion, migration, pattern formation, and signal transduction (Colombatti et al., 1993). Each *AqAF* gene also encodes a single Wreath domain at the 3' end. The Wreath domain has only been found in other sponge AFs and is important for the formation of the sunburst/ring shape (Richardson, 1981; X Fernandez-Busquets & Burger, 2003).

All the *AqAFs* are very polymorphic and possess different levels of polymorphism throughout their sequences between individuals. This variation results in each individual having a

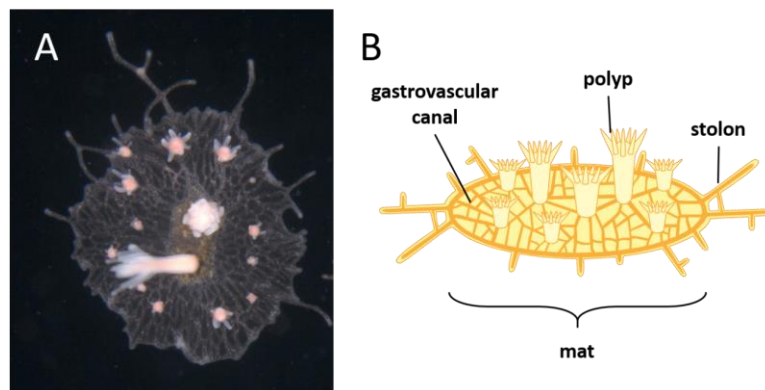
unique AF sequence pattern through which allorecognition can take place. In addition to individual variation, one larvae sequenced showed evidence of nucleotide differences between the genomic and cDNA sequences, suggesting possible RNA editing variability (Grice et al., 2017). This would also increase the uniqueness of AqAF recognition identities even between closely related or sibling sponges. This variation parallels the alloincompatibility observed in sponges and is expected to play a role in allorecognition through cell adhesion. The AFs in sponge allorecognition systems suggest that there may be an evolutionary relationship between cell adhesion and histocompatibility systems previously underappreciated.

### **1.6.3 *Hydractinia symbiolongicarpus***

Cnidarians are among one of the clearest examples in which self/non-self recognition, allorecognition, is observed. While nearly all cnidarians have these genetic systems, very few have been characterized. The only cnidarian model for allorecognition in which genes have been identified is *Hydractinia symbiolongicarpus*. In addition to their long-standing history in research, *Hydractinia* have several advantages over other allorecognition models. They have a well-defined and robust life cycle and are relatively easily and inexpensive to maintain. Decades of research on the phenotypic and genetic elements surrounding allorecognition have laid a firm foundation from which greater understanding can be obtained.

In nature, *Hydractinia* grow on shells inhabited by a hermit crab living primarily in the intertidal zone (Cunningham et al., 1991; Buckley & Ebersole, 1994; Weissberger, 1995; Nicotra & Buss, 2005). *Hydractinia* morphology is relatively simple with four main features consisting of polyps, gastrovascular canals, stolons, and a plate of tissue referred to as the mat (Figure 8A,B). Polyps extend upward from the mat and are connected through a network of gastrovascular canals.

Protrusions of the gastrovascular canals outside the mat area are defined as stolons. Polyps will differentiate into four functional subtypes including feeding polyps (gastrozooids), reproductive polyps (gonozooids), and defense polyps (dactylozooids and tentaculozooids) (W A Mueller, 1964; Sanders et al., 2014). Tentacles protrude from the gastrozooid and contain nematocysts that are used to paralyze and pull prey into the hypostome, or mouth. The food is digested in the body column and then distributed through the colony via the gastrovascular canals, and any waste generated is expelled through the hypostome. Gonozooids develop and house the gametes in compartments called gonophores. Dactylozooids harbor many nematocysts (stinging cells) which are thought to help with food gathering in addition to defense. Tentaculozooids are a less common polyp found on *Hydractinia* which also aid in defense. Neither dactylozooids nor tentaculozooids appear in laboratory-maintained colonies.



**Figure 8. *Hydractinia symbiolongicarpus* morphology.**

A) Colony growing on slide. B) Labeled parts of the colony.

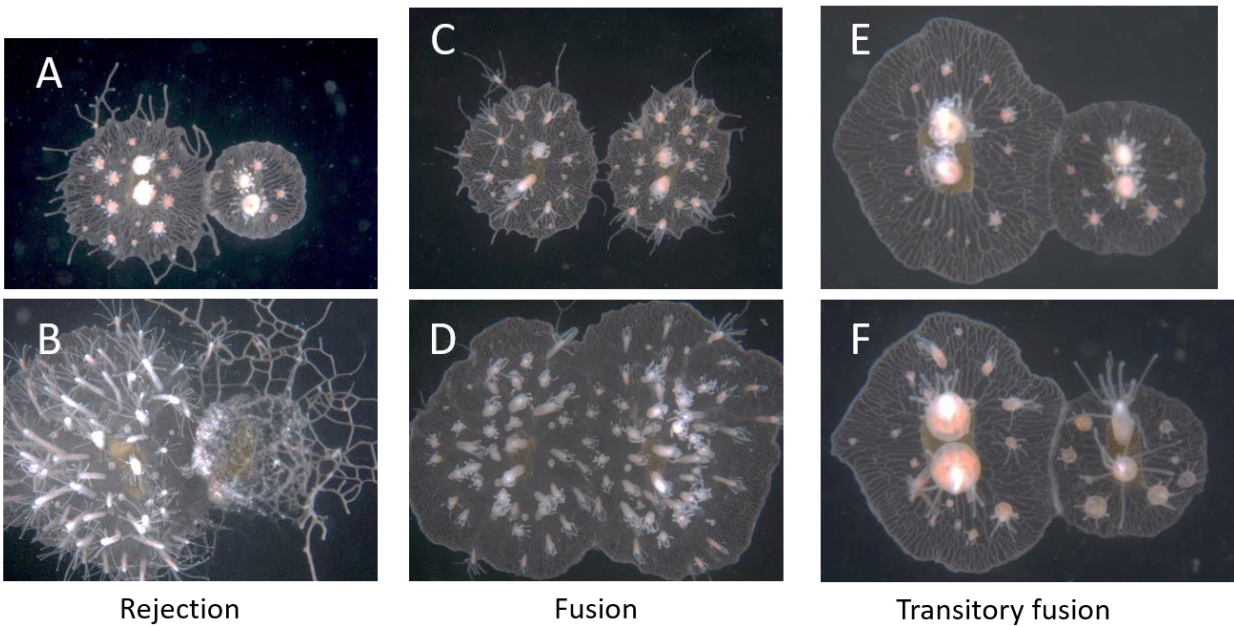
The full life cycle of *Hydractinia* lasts approximately two to three months. However, adult *Hydractinia* have been maintained for at least a decade under laboratory conditions with no visible signs of senescence. *Hydractinia* are dioecious and the gonophores will only produce sperm or egg



based on its stem cells. Gametes are released approximately 1 hour after being exposed to light after a light:dark cycle. Healthy *Hydractinia* can release gametes daily, allowing for reliable material for breeding and experiments. After fertilization, embryos will develop into larvae over 1-4 days and can then metamorphose into primary polyps either through natural cues (Muller, 1973) or through chemical induction (Spindler & Werner, 1972; G Plickert et al., 1988; Blackstone, 1996). The single polyp will expand by extending its stolons and generating mat tissue and additional polyps (Catherine S. McFadden & Buss, 1984). A *Hydractinia* colony will become sexually mature in 2-3 months. *Hydractinia* stem cells, also called I-cells, control regeneration and the gametic output and are circulated throughout the colony as opposed to being housed in one location.

All alloresponses are initiated when the borders of mat tissue or stolons grow into contact between two *Hydractinia* (W A Mueller, 1964; Buss et al., 1984; R. Lange, 1989). The alloresponse is species-specific and does not occur when a *Hydractinia* grows into contact with another species (Buss et al., 1984; Ivker, 2014). Within 2 hours of making contact, nematocytes (stinging cells) migrate to the site of contact (W A Mueller, 1964; Buss et al., 1984). By approximately 2.5 hours post-contact, the genetic compatibility between the colonies will determine whether they continue with a rejection response or exhibit a fusion response (R. Lange, 1989). When colonies are genetically incompatible, the gathered nematocysts will discharge into the opposing colony to kill any foreign tissue (Figure 9A,B). The colonies will continue to shoot their nematocysts until contact is broken or one animal dies. Colonies can also grow their mat tissue over the other to limit resources and take over space. When colonies are compatible, the nematocysts will disperse and the colonies will begin to fuse their gastrovascular canals and mat borders so that resources, including stem cells, can be shared across the expanded colony (Figure

9C,D). When colonies are only partially compatible, they exhibit a “transitory fusion” phenotype, in which they initially fuse and within one to two days form a border which prevents resources being shared between the two colonies (Figure 9E,F). This separation is hypothesized to be a controlled cell death mechanism but is not fully understood yet (Buss et al., 2012).



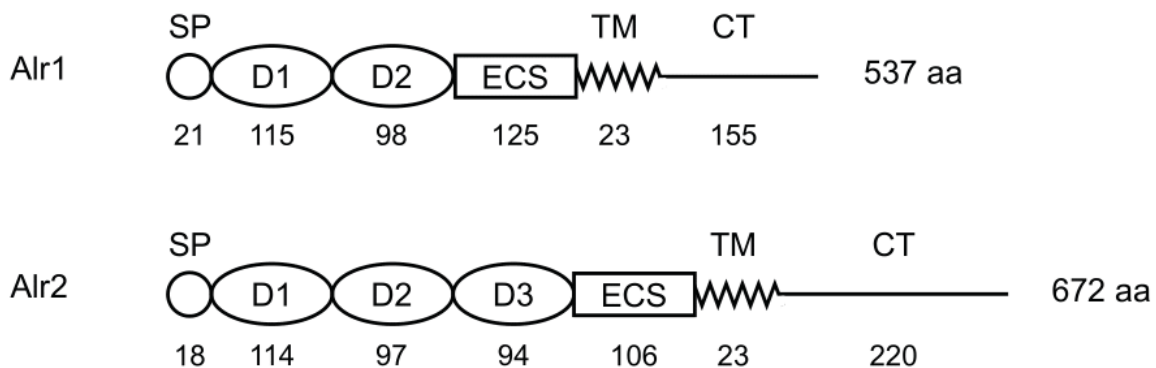
**Figure 9. *Hydractinia* alloresponses.**

A,B) Pre- and post-rejection response. C,D) Pre- and post-fusion response. E) Initial fusion between partially compatible colonies. F) Barrier formed between partially compatible colonies.

Previous experiments with inbred, laboratory strains of *Hydractinia* have demonstrated that colonies can distinguish self from non-self by their genotype at two linked genes called *Allorecognition 1* (*Alr1*) and *Allorecognition 2* (*Alr2*) (Cadavid et al., 2004; Powell et al., 2007, 2011). In general, animals that shared at least one allele at both loci fused, while those that shared no alleles at either *Alr1* or *Alr2* always rejected. If colonies only shared alleles at one locus, they most often underwent transitory fusion. Because only two alleles were present at each locus in

these strains, it was impossible to determine how similar alleles needed to be for colonies to fuse. In some of the experiments using outbred colonies, knowing the identity of *Alr1* and *Alr2* was not always sufficient to predict the allorecognition outcome. Thus, it was hypothesized that at least one additional allorecognition locus must be present in *Hydractinia* (Nicotra et al., 2009; Rosa et al., 2010; Powell et al., 2011).

*Alr1* and *Alr2* both encode type I transmembrane proteins with two or three tandem Ig-like domains in their extracellular regions (Figure 10) (Nicotra et al., 2009; Rosa et al., 2010). There are no known homologs for *Alr1* or *Alr2* based on their full length sequence. However, the individual ectodomains may have structural homology to Ig-like domains based on structural predictions. 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).



**Figure 10. Alr1 and Alr2 domain architecture.**

SP = Signal peptide, D1/D2/D3 = Domain 1/2/3, ECS = Extracellular spacer, TM = Transmembrane domain, CT = Cytoplasmic tail. Number of amino acids in each domain is given below each model. The total amino acid (aa) of each protein is approximate due to slight variability between alleles.

*Alr1* and *Alr2* are both 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 that hundreds of distinct binding specificities could exist in nature. It has been assumed that novel *Alr1* and *Alr2* alleles are generated by random mutations that are then subjected to negative frequency-dependent selection.

### 1.7 Common themes in allorecognition

Of the four questions common to all recognition systems, “what are the genes?” and “what polymorphism exists?” are the most straightforward to answer as they rely primarily on sequence data. Indeed, nearly all systems discussed above have the characteristic polymorphic genes to regulate histocompatibility (Table 1).

**Table 1. Summary of known genes involved in allorecognition.**

<b>Model</b>	<b>Specificity gene(s)?</b>	<b>Polymorphic?</b>	<b>Domain(s)?</b>
<i>M. xanthus</i>	<i>traA</i>	Yes	PA14
<i>P. mirabilis</i>	<i>idsD, idsE</i>	Yes, Yes	Unknown, Unknown
<i>N. crassa</i> (VI)	11+ <i>Het</i> loci	Yes	Unknown
<i>N. crassa</i> (SI)	<i>mat-A/mat-a</i>	No/No	Transcription factors
<i>S. commune</i> (VI)	5+ <i>Het</i> loci	Yes	WD40 domain
<i>S. commune</i> (SI)	A $\alpha$ /A $\beta$ , B $\alpha$ /B $\beta$	Yes, Yes	Transcription factors
<i>Brassicaceae</i>	<i>SCR, SRK</i>	Yes, Yes	Unknown, Unknown
<i>Solanaceae</i>	<i>S-RNase, SLFs</i> (16-20)	Yes, Some	S-RNase, F-box
<i>Papaveraceae</i>	<i>PrsS, PrpS</i>	Yes, Yes	Secreted, Receptor
<i>D. discoideum</i>	<i>tgrC1, tgrB1</i>	Yes, Yes	Ig domains
<i>B. schlosseri</i>	<i>BHF</i>	Yes	Unknown
<i>M. prolifera</i>	MAF	Yes	Unknown, glycosylation
<i>H. symbiolongicarpus</i>	<i>Alr1, Alr2</i>	Yes, Yes	Ig domains, Ig domains

The third question regarding the structural domains involved in the interactions is more challenging to answer as predicting or resolving structures for highly variable proteins can be difficult. Several systems, especially those reliant on cell adhesion, possess proteins that bear domains with structural similarity to the immunoglobulin superfamily domain, but their variability makes detecting homology and predicting based on secondary structure a challenge. As observed in fungi and plants, some systems only use heterophilic interactions that act as ligand-receptor pairs, thereby increasing the number of possible structural domains through which recognition can occur.

The fourth question, “How do new self-identity specificities evolve?”, is perhaps the most difficult to answer. Part of the difficulty in addressing this question is attributed to the lack of comprehensive data in relation to both time and sample size. Many sequences have numerous differences between recognition alleles, which make meaningful conclusions regarding the evolution of self-identity specificities hard to resolve without extensive experimental data.

The studies presented in this dissertation address these four questions in the cnidarian model *Hydractinia*. While *Alr1* and *Alr2* are most definitely involved in determining histocompatibility, data shows that there may be additional allodeterminants or *Alr*-like genes whose roles and degree of polymorphism remain unknown. How *Alr1* and *Alr2* evolve new binding specificities is yet to be determined. The domains of these proteins have been previously predicted based on weak sequence homology but lack empirically derived structures to verify their folds. Recent advances with whole genome sequencing have made it possible to characterize the genomic region containing these genes and to begin addressing these questions.

In Chapter 2, I address the outstanding question of whether additional allorecognition-like genes are present in the ARC. I first show the assembly of the ARC which is located within a large

genomic region of approximately 12 Mb. In addition, I summarize the presence of 40+ distinct *Alr* genes encoded throughout the ARC, which bear similarities to *Alr1* and *Alr2*.

In Chapter 3, I assessed the predicted structural conservation of the *Alrs* identified. I show that all *Alr* genes are predicted to encode individual domains with structural and sequence similarities to known IgSF domains. Furthermore, I show that domain 1 is most similar to a V-set domain, domains 2 and 3 are most similar to I-set domains, and the ECS has a FnIII-like fold.

Through Chapter 4, I address the question of how novel binding specificities evolve in homophilic binding proteins. I show that the evolution of novel binding specificities in homophilic binding proteins can occur through single residue changes in addition to evolving through neutral evolution. Also, I showcase a small clade of *Alr2* domain 1 isoforms that exhibit these phenotypes.

In Chapter 5, I address the sequence and structural variation in the ARC between three haplotypes. I show that the ARC is a region of significant variation in which some *Alr* genes have copy number variation in addition to sequence polymorphism. These observations provide a list of additional allorecognition candidates which may be involved in the alloresponse.

## **2.0 Sequencing and annotation of the Allorecognition Complex yields a 12 Mb region which contains a large Alr family**

### **2.1 Foreword**

This chapter is adapted from a manuscript in submission for publication in which I am first author: Aidan L. Huene, Steven M. Sanders, Zhiwei Ma, Anh-Dao Nguyen, Sergei Koren, Manuel H. Michaca, Jim C. Mullikin, Adam Phillippy, Christine E. Schnitzler, Andreas D. Baxevanis, and Matthew L. Nicotra (2021, Unpublished)

### **2.2 Summary**

In *Hydractinia*, *Alr1* and *Alr2* contribute to allorecognition, but experiments with outbred colonies have suggested that additional allorecognition genes may exist in the genome. To date, no homologs have been identified for either *Alr1* or *Alr2* leaving their evolutionary history and relationships a mystery. Until recently, only limited sequence surrounding *Alr1* and *Alr2* was known to contain several *Alr*-like gene candidates, strengthening the hypothesis that additional allorecognition genes may be involved. To determine the extent of the ARC and whether additional allorecognition-like genes may be present in the genome, the whole genome of *Hydractinia* was sequenced and assembled. The sequence containing *Alr1*, *Alr2*, and the other known markers were used to identify and expand the ARC. This yielded a large genomic region spanning approximately

12 Mb. Through *de novo* prediction and BLAST searches, I was able to annotate 41 distinct *Alr*-like loci.

## 2.3 Introduction

In previous studies, efforts to map allorecognition genes in *Hydractinia* relied on inbred lines, referred to as F and R. Results from these studies suggested that allorecognition segregated as a single Mendelian codominant expression (Mokady & Buss, 1996). Further study revealed recombination in defined genetic lines and yielded two linked loci which control allorecognition (Cadavid et al., 2004) that have differential effects on the allorecognition phenotype (Powell et al., 2007). This latter work generated a partial sequence for the ARC from the F haplotype and was named ARC-F. From the ARC-F sequence, two genes were positionally cloned and identified to control allorecognition in inbred lines, *Allorecognition 1* (*Alr1*) and *Allorecognition 2* (*Alr2*) (Nicotra et al., 2009; Rosa et al., 2010). To date, no homologs have been identified for either *Alr1* or *Alr2* in other organisms.

Several lines of evidence suggested the *Hydractinia* ARC contains one or more additional allorecognition genes. First, a colony's genotype at *Alr1* and *Alr2* can fail to predict allorecognition responses. This occurs infrequently between strains inbred for their fusibility (Cadavid et al., 2004; Powell et al., 2007) and more frequently between outbred colonies selected from the wild (Nicotra et al., 2009; Rosa et al., 2010). Second, the genomic region surrounding *Alr1* contained an array of open reading frames encoding peptides similar to the IgSF-like domains of *Alr1*. Similarly, the genomic region immediately upstream *Alr2* contains several *Alr2* pseudogenes (Nicotra et al., 2009; Rosengarten et al., 2011).

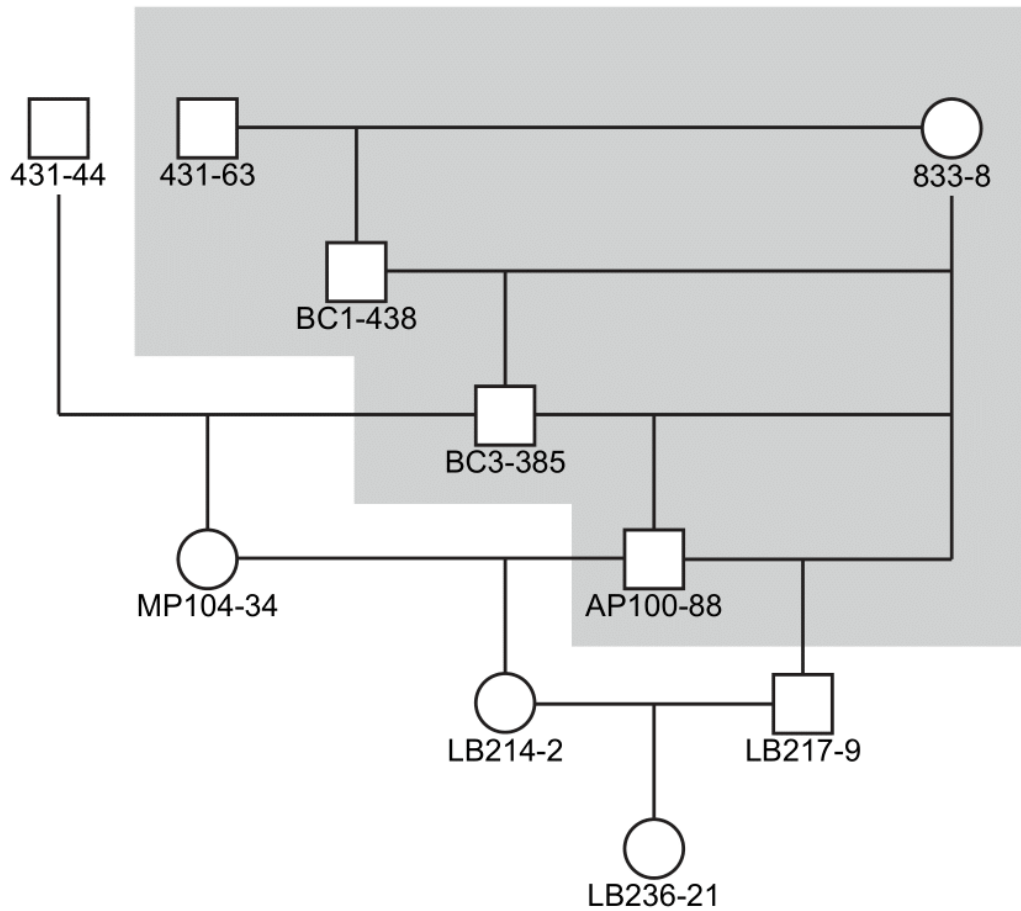


Together, these observations have led to the hypothesis that the ARC contains more allodeterminants. The primary candidates are the *Alr*-like sequences that have been identified so far, plus any that might exist in the regions that have yet to be sequenced. With the recent sequencing of the *Hydractinia* genome, it is now possible to search for additional candidate genes which may share sequence homology to *Alr1* and *Alr2*. Additional *Alr*-like sequences may also provide additional data to search for homologs in other species. In addition, identifying any other related genes may elucidate potential ancestry of these genes in the invertebrate allorecognition system. The diversity present in both *Alr1* and *Alr2* may also be important for the evolution of diversity in maintaining the allorecognition system in *Hydractinia*. Here, I report the assembly and annotation of a nearly complete reference sequence for one ARC haplotype. I identified a family of 41 *Alr* genes, putative genes, and pseudogenes.

## **2.4 Results**

### **2.4.1 Assembly of ARC reference sequence**

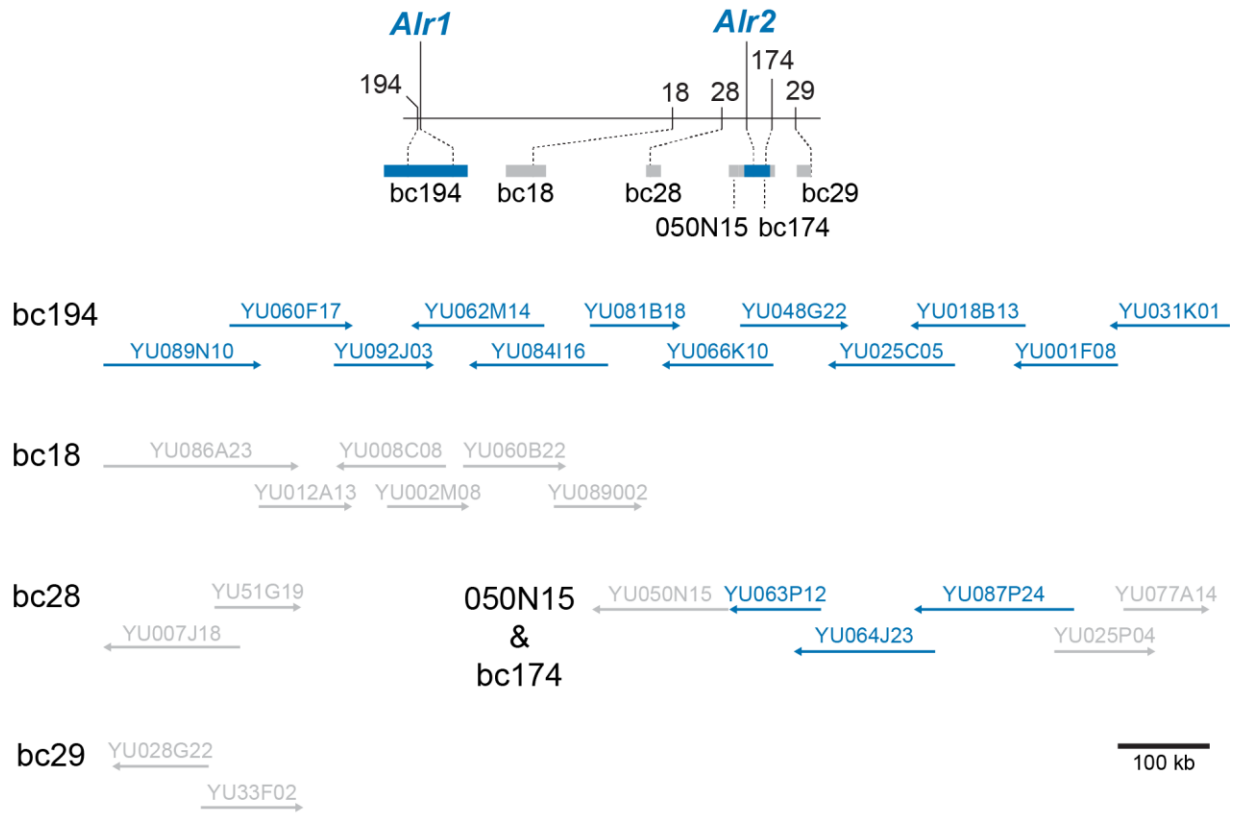
In previous work, a partial sequence for the ARC-F was generated by creating a BAC library from an ARC-F homozygote (colony 833-8 in Figure 11) and subsequently performing five chromosome walks, each starting from a marker in the ARC linkage map (Figure 12) (Powell et al., 2007; Nicotra et al., 2009; Rosa et al., 2010).



**Figure 11. Pedigree of colonies used to generate ARC-F reference sequence.**

The pedigree of colony LB236-21 can be recreated by concatenating previously published pedigrees (shaded area) (Cadavid et al., 2004; Powell et al., 2007). Colony AP100-88 is from the mapping population in Powell et al. (2007).

Colony 431-44 is from the mapping population in Cadavid et al. (2004).



**Figure 12. Detail of the initial ARC reference assembly.**

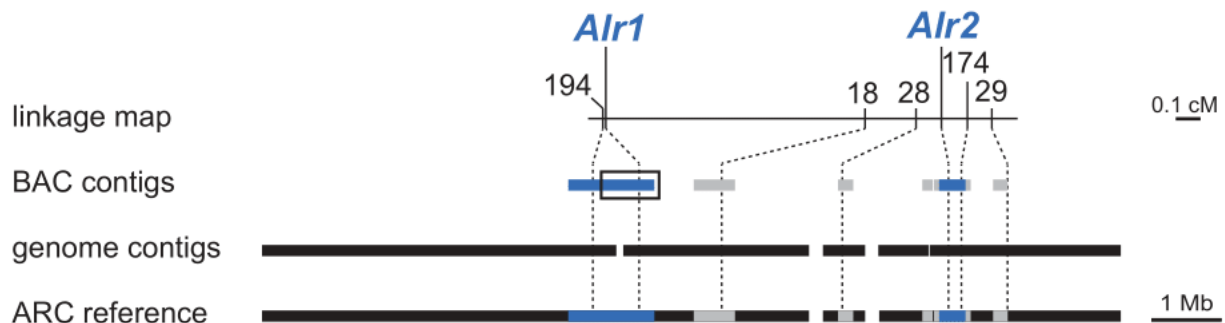
A) Minimum tiling path of sequenced BAC clones resulting from chromosome walks from five markers in the ARC linkage map. Clone names are indicated above an arrow indicating their orientation. Sequences reported in Rosa et al (2010) or Nicotra et al (2009) are in navy blue. Unpublished sequences are in gray.

The minimum tiling path of each walk was sequenced, resulting in six BAC contigs with a total length of 2.9 Mb (Figure 12). *Alr1* and *Alr2* were then identified via positional cloning (Nicotra et al., 2009; Rosa et al., 2010). The genomic region surrounding *Alr1* contained an array of open reading frames encoding peptides similar to the IgSF-like domains of *Alr1* (Rosa et al., 2010).

To determine the full extent of the *Alr* gene family, I sought to generate a complete reference sequence for the ARC-F haplotype. The genome was sequenced and assembled of a

second ARC-F homozygote, colony 236-21, which is a descendant of colony 833-8 (Figure 11). High-molecular weight genomic DNA was sequenced via PacBio long-read sequencing and polished with Illumina data to create a high-quality genome assembly. The resulting non-filtered assembly was 431 Mb long, with 5697 contigs and an N50 of 397 Kb.

To find contigs that would align with and extend the preexisting ARC-F sequence, I aligned the original BAC contigs to this new genome assembly using NUCmer (Delcher et al., 2002; Kurtz et al., 2004). I identified five genome contigs that overlapped the BAC contigs with >99% sequence identity (Figure 13, Table 2). The only major discrepancies between the BAC contigs and the genome contigs were in repeat regions. Therefore, I merged these sequences by filling the gaps between BAC contigs with sequences from the genome assembly. The resulting ARC-F reference sequence spanned 11.83 Mb and contained two gaps of unknown size (Figure 13).



**Figure 13. Assembly of the *Alr* gene complex.**

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).

**Table 2. Overlap coordinates of genomic contigs and BAC contigs used to create reference ARC-F sequence.**

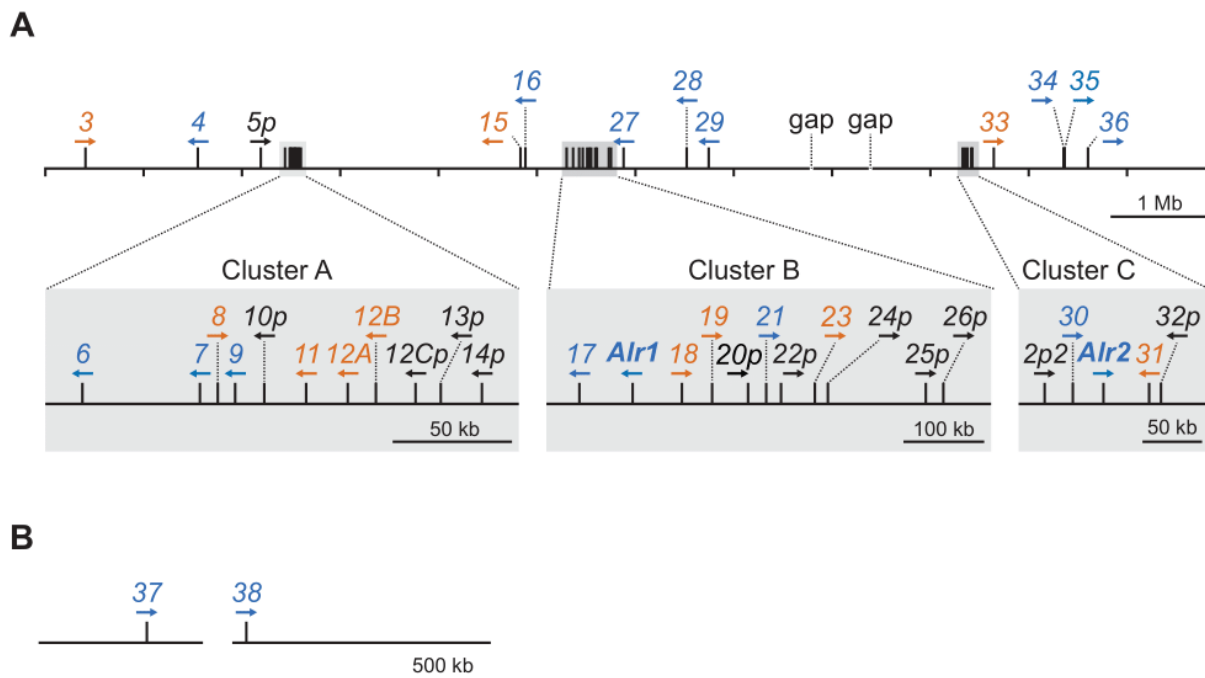
Genome Assembly			BAC contigs		
ID (length)	start	stop	ID (length)	start	stop
utg0000000001 (5,049,836 bp)	687,083	1	bc194 (1,225,536 bp)	1	684,144
utg0000000021 (2,644,760 bp)	174	442,144	bc194 (1,225,536 bp)	783,992	1,225,536
utg0000000021 (2,644,760 bp)	1,005,148	1,590,913	bc18 (586,384 bp)	1	586,384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

#### 2.4.2 The *Hydractinia* genome contains a large family of *Alr* genes and pseudogenes

Next, I annotated all *Alr*-like genes in the reference sequence. To do this, I generated *ab initio* gene predictions with AUGUSTUS (Stanke et al., 2004), mapped RNAseq data from colony 236-21 to the ARC-F reference sequence, and used BLASTX to identify regions predicted to encode proteins similar to *Alr1* and *Alr2*. These data were loaded into the Apollo annotation platform (Dunn et al., 2019) where I created gene models. As new gene models were created, I used them in iterative TBLASTX searches to identify *Alr* genes that might not have been detected via similarity to *Alr1* or *Alr2*. Finally, to identify *Alr* genes that might exist outside the ARC-F reference sequence, I used TBLASTX to query the full genome assembly with the amino acid translation of each *Alr* gene model. All *Alr* genes and pseudogenes were 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 assigned the same number followed by a letter (e.g., *Alr12A* and *Alr12B*).

In total, I created 41 gene models (Figure 14A). Almost all (39/41) were located in the ARC reference sequence. More than half (27/41) were encoded within one of three *Alr* clusters that I named A, B, and C (Figure 14A). The remaining two genes, *Alr37* and *Alr38*, were each on separate contigs not contiguous with the ARC-F reference sequence (Figure 14B).

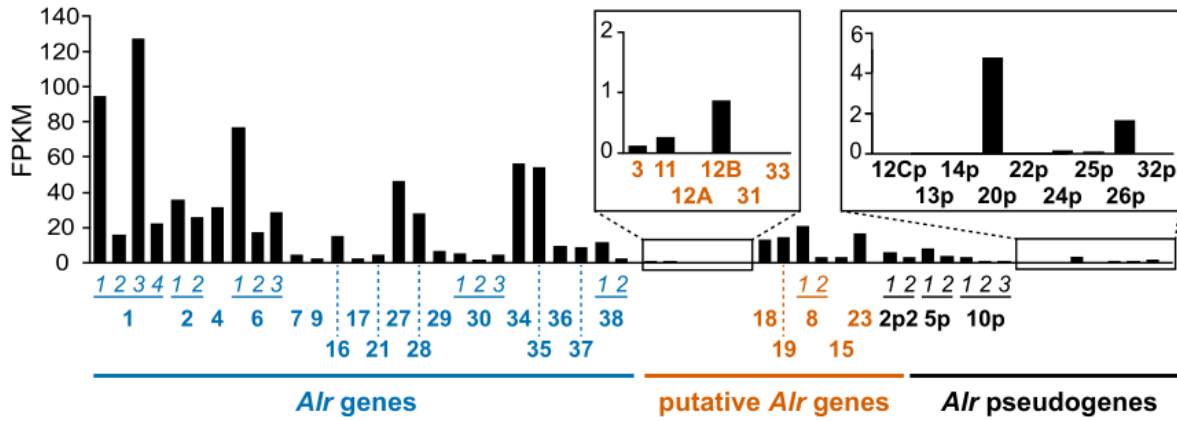


**Figure 14. Annotation of *Alr*-like genes in the Allorecognition Complex.**

A) Identity, location, and orientation of *Alr* family members within the ARC reference (blue, bona fide gene; orange, putative gene; black, pseudogene). B) Two *Alr* genes located in genome contigs that could not be physically linked to the ARC reference sequence.

To estimate the expression level of each gene model, I calculated the fragments per kilobase of transcript per million mapped fragments (FKPM) from the RNAseq data. Gene models

with less than 1 FPKM were deemed unexpressed. The results suggested that *Alr1*, *Alr2*, *Alr4*, *Alr6*, *Alr27*, *Alr28*, *Alr34*, and *Alr35* are most highly expressed (Figure 15).



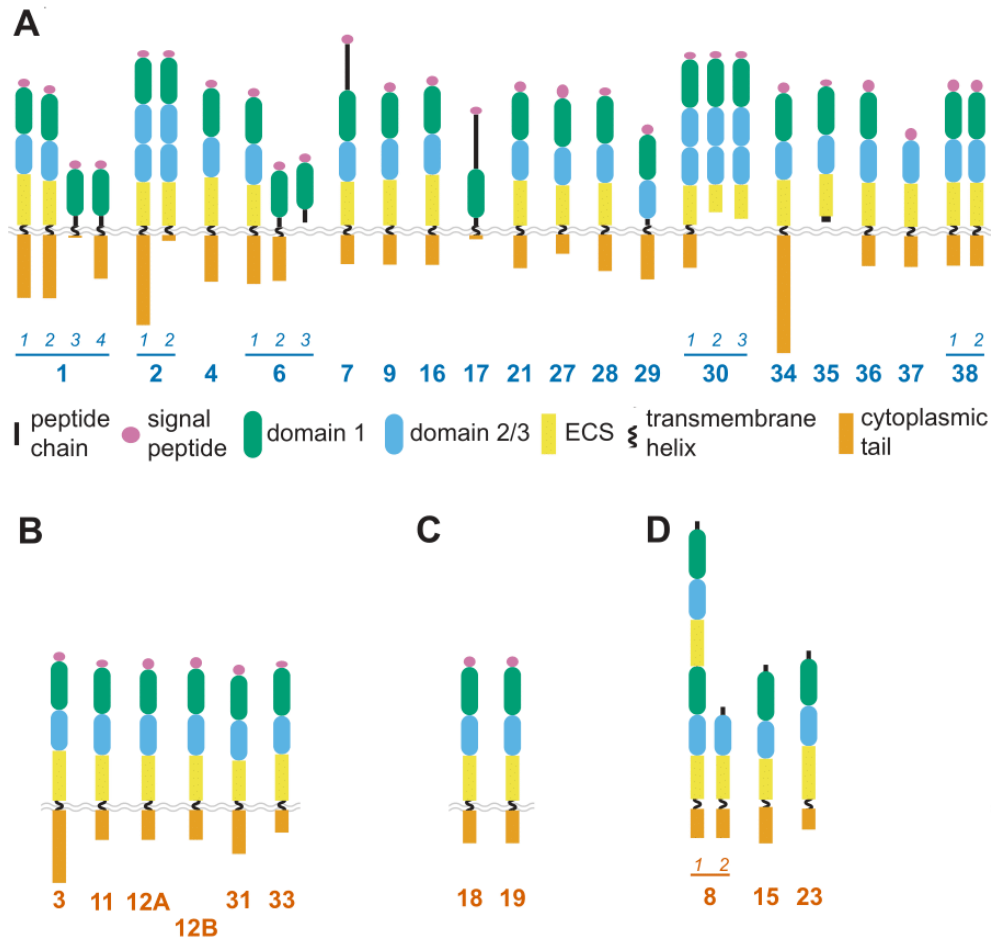
**Figure 15. Expression of *Alr* genes.**

Estimated expression levels of each *Alr* gene, putative gene, and pseudogene. FPKM, fragments per kb mapped.

Genes are identified by bold numbers. Splice variants are indicated by horizontal lines and numbers in italics.

Gene models varied in the amount of RNAseq data supporting each exon, the size of their predicted open reading frames (ORFs), and the domain architecture of their gene products. Therefore, to aid my annotation, I classified them as either a bona fide gene, a putative gene, or a pseudogene based on the following criteria.

A gene model was classified as a bona fide gene if it had a single ORF and the RNAseq covered every exon with evidence for proper splicing. Eighteen gene models fit this definition, including *Alr1* and *Alr2* (Figure 16A). As shown in Figure 16A, the *Alr* genes generally encode single-pass transmembrane proteins with 1-3 tandem extracellular domains, an ECS, a transmembrane helix, and a cytoplasmic tail. Without exception, each extracellular domain, ECS, and transmembrane helix was encoded by a single exon.



**Figure 16. Domain architecture of *Alr* genes.**

A) Domain architecture of putative *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. B) Domain architecture of unexpressed putative genes. C) Domain architecture of expressed putative genes. D) Domain architecture of putative genes lacking a predicted signal peptide.

A gene model was classified as a putative gene if lacked some of the clear RNAseq features of a bona fide gene but was not obviously a pseudogene. Eleven gene models fit this definition (Figure 16B-D). Six had clear sequence similarity to bona fide *Alr* genes but were unexpressed (Figure 15, Figure 16B). I did not call these pseudogenes as they may be expressed at developmental time points or tissues not represented in the RNAseq dataset. Two gene models had



ORFs that would encode a full Alr protein. However, there was no evidence of splicing between exons 1, 2, and 3 (Figure 16C). Three gene models were not predicted to encode a signal peptide (Figure 16D), which would give them an inverted membrane topology relative to other Alr proteins.

A gene model was classified as a pseudogene if it had sequence similarity to a bona fide or putative *Alr* gene but was truncated by nonsense or frameshift mutations. Eleven gene models fit this definition (Table 3). Several pseudogenes were expressed at modest levels relative to other *Alr* genes (Figure 15).

**Table 3. Gene models classified as *Alr* pseudogenes.**

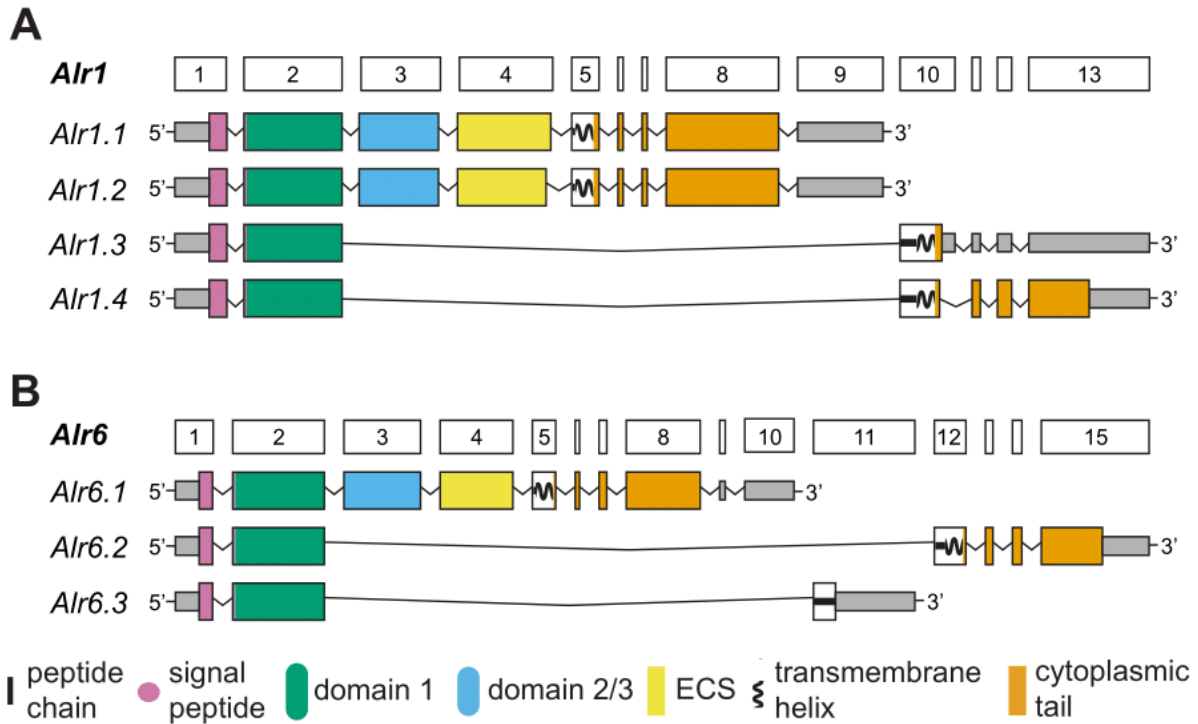
<b>Gene model name</b>	<b>Expression</b>	<b>Reason for classifying as a pseudogene</b>
Alr2p2.1	yes	Partial duplication of exons 1-4 of Alr2. Frame-shift in exon 3 leading to premature stop codons in exon 3.
Alr2p2.2	yes	Partial duplication of exons 1-5 of Alr2. Frame-shift in exon 3 leading to premature stop codons in exon 3.
Alr5p	yes	No evidence of splicing between exons 2-3.
Alr10	yes	Improper splicing of exon 4 to downstream exons introduces stop codon in transcript.
Alr12C	no	Stop codon in exon 2.
Alr13	no	Stop codon in exon 1.
Alr14p	no	No evidence of expression or exon encoding signal peptide.
Alr20p	yes	No evidence of expression or splicing between exons 1 and 2. No evidence of exon encoding a signal peptide.
Alr22p	no	No evidence of expression or exon encoding signal peptide.
Alr24p	yes	A few reads map to exons 2-3. No evidence of exon encoding a signal peptide.
Alr25p	yes	A few reads map to exons 2-5. No evidence of exon encoding a signal peptide.
Alr26p	yes	No open reading frame. No evidence of exon encoding a signal peptide. No splicing between exon 2-3.
Alr32p	no	Only three exons, which have sequence similarity to Alr31, but are not expressed.

In particular, I paid attention to the region directly upstream of *Alr2* because it contained two gene models that had been identified as pseudogenes in previous publications. The first, immediately upstream of *Alr2*, was named CDS6P by Nicotra et al. (2009) and *alr2P1* by Rosengarten et al. (2011). It comprised four exons similar to the first four exons of *Alr2* and had been assumed to be a non-functional partial duplication. Here, I was able to identify additional exons that encoded a transmembrane domain, cytoplasmic tail, and 3' UTR. These new exons were not homologous to *Alr2*. As described above, I therefore classified this locus as a gene and named it *Alr30*. The second gene model had also been thought to be a duplication of the first four exons of *Alr2* and was called CDS5P by Nicotra et al. (2009) and *alr2P2* by Rosengarten et al. (2011). Here, I also concluded the locus was a pseudogene. For consistency with previous work, I have named this pseudogene *Alr2p2*.

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

Several *Alr* genes were alternatively spliced in ways that would change the domain architecture of the encoded protein. At *Alr1*, for example, I identified four splice variants, including two that had been previously reported (*Alr1.1* and *Alr1.2*) (Rosa et al., 2010). In *Alr1.3* and *Alr1.4*, exon 2 was spliced to exons 10-13, which are new exons that were not reported by Rosa et al. (2010) (Figure 17A). *Alr1.3* and *Alr1.4* had different splice donors in exon 10, introducing a frameshift in *Alr1.3* and causing it to have a shorter cytoplasmic tail (Figure 16A). Alternative splicing was also observed at *Alr6* (Figure 16A, Figure 17B). *Alr6.1* encoded a protein with two IgSF-like domains, followed by an ECS, TM, and cytoplasmic tail. However, in *Alr6.2*, exon 2 was spliced to exon 12 to generate a protein with an N-terminal IgSF-like domain followed by a

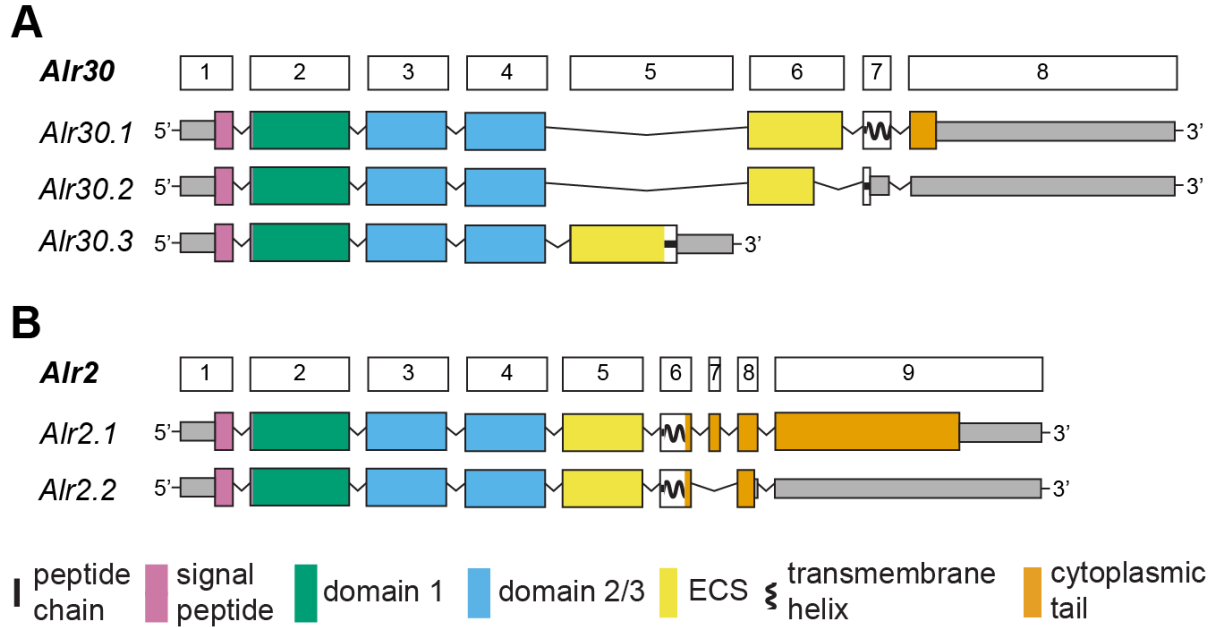
different TM and cytoplasmic tail. In *Alr6.3*, exon 2 was spliced to exon 11 and lacked a TM, raising the possibility that its gene product is secreted.



**Figure 17. Alternative splicing of *Alr1* and *Alr6*.**

A) Alternative splicing of *Alr1*. B) Alternative splicing of *Alr6*. In (A) and (B), exons are colored according to the type of domain/region they encode.

A similar splicing pattern, potentially leading to secreted gene products, was observed in *Alr30* and *Alr35* (Figure 15 and Figure 18A). At *Alr2*, the splicing pattern of multiple reads indicated the presence of transcripts lacking the 22-bp exon 7, which would introduce a frameshift that truncates the cytoplasmic tail (Figure 15 and Figure 18B).

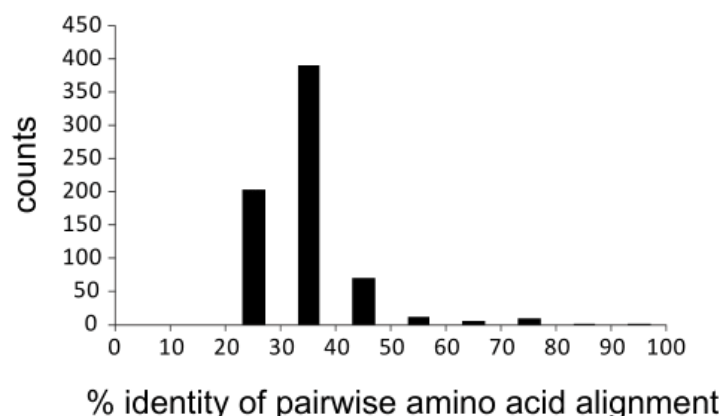


**Figure 18. Alternative splicing at *Alr30* and *Alr2*.**

Alternative splicing of *Alr30* (A) and *Alr2* (B). Exons are colored by the type of region they encode.

#### 2.4.4 Sequences of *Alr* family members are highly diverse

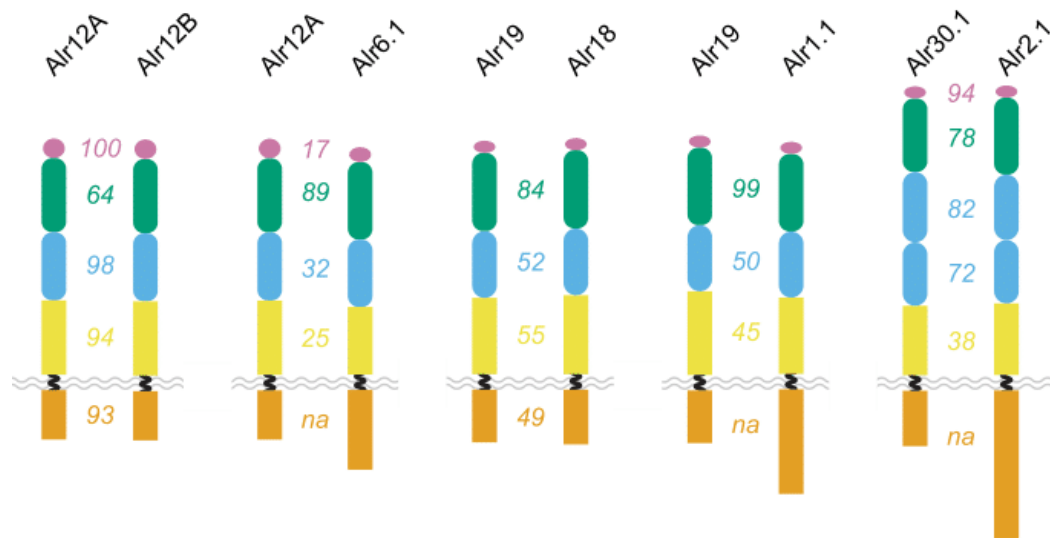
To investigate the evolutionary relationships between *Alr* family members, I attempted to create a single multiple sequence alignment of the predicted amino acid translations of all *Alr* genes and putative genes. However, their sequences were so divergent that it was impossible to obtain a high-quality alignment even after restricting the alignment to sequences of similar length. This drove us to assess overall sequence similarity within the family by performing all possible pairwise alignments. I found that the average percent identity between any two *Alr* protein sequences (excluding splice variants of the same gene) was  $24.3\% \pm 8.6\%$  (Figure 19). Only 2% of pairwise alignments had more than 50% identities. Thus, a substantial amount of sequence evolution has occurred since the origin of the *Alr* family.



**Figure 19. Pairwise amino acid alignment between Alr extracellular domains.**

Histogram of amino acid percent identities for all possible pairwise alignments of Alr genes and putative genes.

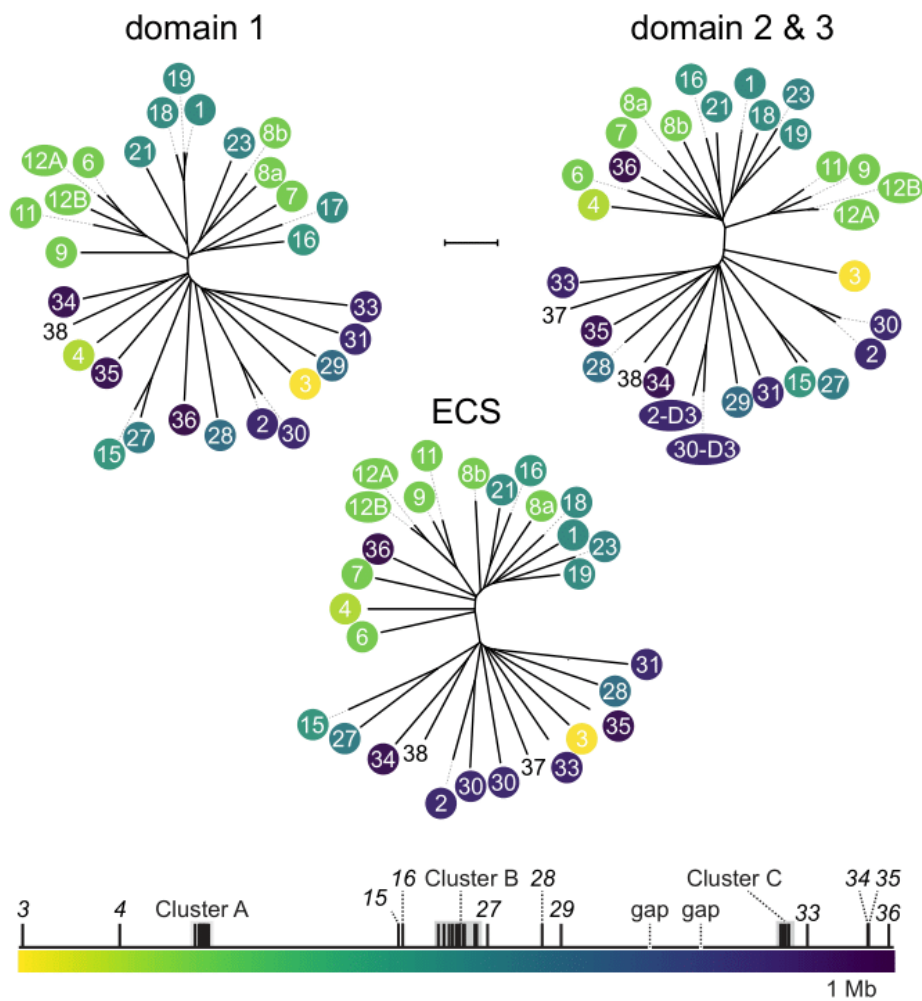
Closer analysis of the pairwise alignments suggested a history of exon shuffling between *Alr* genes. Specifically, I found several cases in which two predicted proteins were very similar in some domains but very divergent in others. For example, *Alr12A* and *Alr12B* were >90% identical along their entire length except for domain 1, which was only 64% identical (Figure 20). Similarly, *Alr6.1* and *Alr12A* were 89% identical over domain 1, but the remainder of the extracellular region was <32% identical and the cytoplasmic tails were unalignable (Figure 20). Comparable alignments were also found between *Alr18* and *Alr19*, *Alr1.1* and *Alr19*, *Alr2.1* and *Alr30.1* (Figure 20). 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). The data suggests exon shuffling could be a common feature of *Alr* genes.



**Figure 20. Sequence similarity between Alr extracellular domains.**

Evidence of exon shuffling between *Alr* genes. Numbers between protein pairs indicate sequence identity within that domain/region. “na” indicates a region that could not be aligned.

These patterns led us to compare the same domains from different Alr proteins. Therefore, I subdivided each sequence into its constitutive extracellular domains. Then, I produced multiple sequence alignments and neighbor-joining trees (Figure 21) for each domain type. This revealed a pattern in which domains were more similar if they were encoded near each other in the genome. 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. It also suggests exon shuffling does not occur frequently between Cluster C and Clusters A/B.



**Figure 21. 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.

## 2.5 Discussion

The ARC was originally described as a two-locus linkage group, with unknown physical size and gene content (Cadavid et al., 2004). Here, I show that it is encoded within at least 11.83 Mb and contains a large family of genes homologous to *Alr1* and *Alr2*. I have named these

sequences the *Alr* gene family, being mindful of the fact that the potential role of these new *Alr* genes in allorecognition responses remains unknown.

The additional members of the *Alr* gene family previously unknown could explain the appearance of unexpected allorecognition responses in some colonies. The ARC was originally delineated by mapping sequence tagged sites to allorecognition phenotypes in inbred lines (Cadavid et al., 2004). The two outermost sequence tagged sites (markers 194 and 29 in Figure 2A) defined a genomic region containing all polymorphisms that affected allorecognition in these defined genetic lines (Powell et al., 2007, 2011). As shown in this chapter, I now know that the *Alr* gene family extends beyond this region. Specifically, it includes *Alr3-Alr5*, Cluster B, and *Alr33-Alr38*. Some or all of these genes may have been rendered homozygous in the inbred strains. If so, this would have prevented the identification of any additional allodeterminants. An outbred colony's genotype at *Alr1* and *Alr2* would therefore be unable to fully predict allorecognition phenotypes because it would ignore important variation at other allorecognition genes. Variation in these unknown genes would probably skew results toward unexpected rejections instead of transitory fusions or transitory fusions instead of fusions as was observed in prior studies (Nicotra et al., 2009; Rosa et al., 2010). However, this does not account for the infrequent appearance of unexpected phenotypes within inbred strains (Cadavid et al., 2004; Powell et al., 2007). In this case, gene conversion, non-allelic homologous recombination, or unequal crossing over, perhaps promoted by the genomic structure of the ARC, might have generated mutations in *Alr1* or *Alr2*. Further studies with these lines, including fine-scale mapping and genotyping, would ultimately resolve this issue.

The decision to classify some *Alr* sequences as putative genes was primarily based on whether the genes were expressed and correctly spliced. While I have high confidence in the



RNAseq collected and analyzed, there are two caveats associated with this expression data. First, the RNAseq data was primarily intended to guide annotation, and thus did not include biological or technical replicates. The resulting expression levels reported here 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 I as well as others have been consistently unable to isolate high quality RNA from these tissues (unpublished data; Uri Frank, personal communication). Some putative genes may be expressed exclusively in these tissues or at developmental time points not represented in the dataset. Therefore, I expect this gene classification system to be refined as additional expression data is generated.

The RNAseq data does show clear evidence for alternative splicing in several *Alr* genes. Alternative splicing at *Alr1* and *Alr6* leads to shorter peptides with alternative cytoplasmic tails, which might be expected to have different functions. *Alr6* and *Alr30* also have splice variants that lack a transmembrane helix. These might encode secreted peptides. The same may also be true for the *Alr35*, which has only one splice variant that similarly lacks a transmembrane region.

The fact that *Alr* genes have a common domain architecture and are found in tightly linked clusters suggests a history of gene duplication. The low sequence identity between *Alr* sequences, even within clusters, suggests these duplication events are relatively ancient, that the genes have undergone substantial sequence evolution, or both. Resolving the evolutionary history of this family will require the sequencing and analysis of additional ARC haplotypes from *H. symbiolongicarpus* and related hydroids.

## 2.6 Methods

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

Colony 236-21 is a female *Hydractinia* colony and was maintained on glass microscope slides in 38-L aquaria filled with artificial seawater (Reef Crystals) as previously described (Sanders et al., 2018). Colonies were starved for 3 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 mortar, where it froze. The frozen UEB1-tissue mixture was then ground into a fine powder and transferred to a 50 ml centrifuge tube containing additional, 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 repeated inversion. This was centrifuged for 10 min at 3000 x g. The aqueous layer was then 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% ethanol and twice with 100% ethanol. The precipitated material was then gently brought to the bottom of the tube by briefly centrifuging, then 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 the addition of RNAses (RNase cocktail, Ambion, #AM2286 ) and incubation at 37°C for 15 min. 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 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.

Sequencing was performed at the NIH. The genomic DNA was sequenced using 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. Sequences were assembled with the Celera Assembler version 8.3r2 (Berlin et al., 2015) and polished with PacBio reads using the ArrowGrid parallel wrapper (Chin et al., 2013) followed by polishing with Illumina short read data using PilonGrid parallel wrapper (Walker et al., 2014).

## **2.6.2 RNA extraction and sequencing**

RNA was extracted from approximately 30 polyps of colony MN236-21. Polyps were collected by excising polyps directly from the colony using a scalpel, moved to a microcentrifuge tube, briefly centrifuged, and the 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. The lysate was incubated for <5 min at room temperature (RT). Chloroform (100 µL) was then added to the sample and the tube was shaken vigorously for 15 s, followed by a 3 min incubation at RT. The sample was then centrifuged at 12,000 x g for 15 min at 4°C. The aqueous phase was removed and subjected to the PureLink RNA Mini Kit (Invitrogen) per the recommended protocol. RNA quality and quantitation were assessed by Tapestation and Qubit, respectively, at the University of Pittsburgh Genomics Core. Final sample was frozen and stored at -80°C until sequenced.

RNA-seq data was mapped to the genome using HISAT2 (Kim et al., 2015) through the Galaxy public server <https://usegalaxy.org/> (Afgan et al., 2018). Reads were mapped to the genome

under three different parameter settings that were defined as relaxed, normal, and strict. The relaxed settings allowed 10 primary alignments per read and used default paired-end options. Normal settings allowed 5 primary alignments per read and disabled alignments of individual mates. Strict settings allowed only 1 primary alignment per read and disabled alignments of individual mates. Transcript abundance of the *Alr* genes was estimated with cufflinks (Trapnell et al., 2010). Cufflinks was run to quantitate against a reference annotations of the *Alr* genes and correct for multiple read mappings.

### **2.6.3 Assembly of the ARC**

To assemble the full reference sequence for the ARC, NUCmer from the MUMmer package (v3.23) was used to align the BAC contigs with the newly assembled whole genome sequence to identify the contigs that matched the known ARC sequence (Delcher et al., 2002; 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 show only matching hits in one direction on the strands (`-r`) and to remove all hits with 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 & Kucherov, 2005). The genome assembly and BAC sequences were then merged to create a reference sequence of the ARC-F haplotype.

#### 2.6.4 Annotation of *Alr* genes

After obtaining the whole-genome data, genes were predicted using AUGUSTUS (Stanke et al., 2004). 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/>). The masked DNA sequences were then divided into 32 kb segments with 2 kb overlaps. These segments were submitted as BLASTX queries against a database of Alr1 and Alr2 proteins (to identify Alr-like sequences), and the Swiss-Prot database (to identify highly conserved genes). The BLAST results were then concatenated and a custom Perl script was used to adjust their coordinates to align with the unsegmented genomic 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 one primary alignment. The output file (.bam) was then uploaded into Apollo for visualization during annotation. *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.

### **3.0 The *Alr* gene family encodes domains that are novel members of IgSF proteins**

#### **3.1 Foreword**

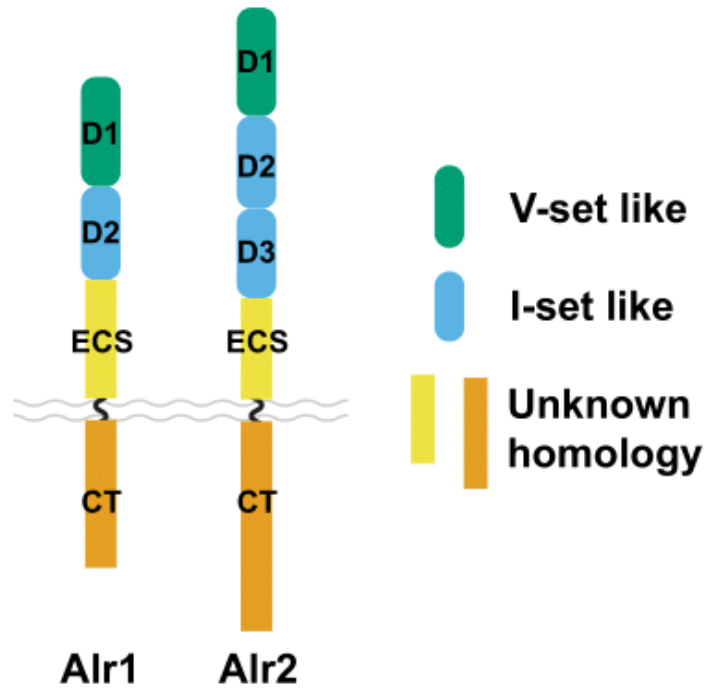
This chapter is adapted from a manuscript in submission for publication in which I am first author: Aidan L. Huene, Steven M. Sanders, Zhiwei Ma, Anh-Dao Nguyen, Sergei Koren, Manuel H. Michaca, Jim C. Mullikin, Adam Phillippy, Christine E. Schnitzler, Andreas D. Baxevanis, and Matthew L. Nicotra (2021, Unpublished)

#### **3.2 Summary**

With the annotation of the ARC complete, the 41 *Alr*-like gene models all shared similar domain architecture suggesting that these genes could have arisen through gene duplication or other mechanisms that resulted in their similarities. To determine what the domains of these genes encode and whether they share any structural similarities between genes, I compared these *Alr*-like genes using sequence analyses and structural predictions. I show that all *Alr* genes are predicted to encode individual domains many of which have structural and sequence similarities to known IgSF domains. Here I show that domain 1 is highly likely to fold as a V-set domain, domains 2 and 3 as I-set domains, and the ECS with a FnIII-like fold.

### 3.3 Introduction

The newly sequenced and annotated ARC revealed additional members of the *Alr* gene family. While the domain architecture of the genes is similar, it was unknown whether these genes shared sequence or structural homology to any known domains. The two previously identified allorecognition genes, *Alr1* and *Alr2*, are known to encode single-pass transmembrane proteins with extracellular domains had similarities to immunoglobulin-like (Ig) domains (Figure 22). These are followed by a region, with no predicted structural folds, called the extracellular spacer (ECS), a transmembrane helix, and a 155 amino acid (aa) cytoplasmic tail for *Alr1* and a longer (220 aa) cytoplasmic tail for *Alr2*. Although clearly related, *Alr1* and *Alr2* have low global sequence identity (<20%), consistent with an ancient gene duplication, rapid sequence evolution, or both. Both proteins also have several tyrosine residues in their cytoplasmic tails, suggesting they might function as phosphorylation-dependent receptors.



**Figure 22. Alr1 and Alr2 domain architecture.**

The putative Ig domain architecture of Alr proteins suggests they could play roles in extracellular protein-protein interactions, signaling, or adhesion. Tandem Ig domains are commonly found in cell adhesion molecules, proteins involved in cell-to-cell communication, and immune receptors. An adhesive function would be consistent with what was already described for Alr1 and Alr2 (Karadge et al., 2015). In addition, the shared domain architecture of Alr1 and Alr2, as well as in the rest of the Alr family, suggests a history of gene duplication.

To further explore this extensive *Alr* gene family, I used sequence and predicted structural homology to study their evolutionary history. Here I report that the vast majority of the *Alr* gene family encode single-pass transmembrane proteins with extracellular Ig domains and, unexpectedly, a fibronectin III (Fn3)-like fold in the ECS. Several Alr proteins also have immunoreceptor tyrosine-based activation motifs (ITAMs) or immunoreceptor tyrosine-based



inhibitory motifs (ITIMs) in their cytoplasmic tails. Invariant cysteines in their extracellular domains may form disulfide bonds within and between domain 2 or 3 and the ECS. Together, these findings mark the discovery of a novel family of immunoglobulin superfamily (IgSF) proteins and provide a slate of candidates for additional genes involved in allorecognition in *Hydractinia*.

### 3.4 Results

#### 3.4.1 Domain 1 is an Ig domain most similar to the V-set family.

Domain 1 of *Alr1* and *Alr2* was originally described as Ig-like because it had some sequence similarity to V-set Ig domains (Nicotra et al., 2009; Rosa et al., 2010). To determine whether domain 1 was V-set-like in the newly discovered Alr proteins, I first used HMMER to compare each sequence to Pfam, a database of hidden Markov models representing known protein families (El-Gebali et al., 2019). At an E-value cutoff of  $<0.01$ , only 6/29 domain 1 sequences were similar to V-set immunoglobulin domains, while a seventh was identified as a family of unknown function (Table 4). To search for more distant homologies, we used HHpred, a method that is more sensitive to remote homologies because it relies on pairwise alignment of hidden Markov models (Zimmermann et al., 2018). I used HHpred to compare each domain 1 sequence to the Structural Classification of Proteins extended (SCOPe), a database that classifies domains from proteins of known structure according to their structural and evolutionary relationships (Fox et al., 2014; Chandonia et al., 2019). Using HHpred, I found that 19/29 sequences had a  $>95\%$  probability of homology with the V-set family of Ig domains (Table 4). These results suggested most domain 1 sequences were remote homologs of V-set domains.

**Table 4. Sequence homology of Domain 1.**

Protein <sup>a</sup>	Domain	HMMER search of pfam			HHpred search of SCOPe		
		Accession	description	e-value <sup>b</sup>	Accession	SCOPe Family	probability <sup>c</sup>
Alr1	D1	PF07686.17	V-set	0.0012	d5l21b	b.1.1.1: V set domains	96.4
Alr2	D1				d5e56a	b.1.1.1: V set domains	95.8
Alr3	D1						
Alr4	D1				d5my6b1	b.1.1.1: V set domains	96.5
Alr6	D1				d5my6b1	b.1.1.1: V set domains	96.5
Alr7	D1				d5l21b	b.1.1.1: V set domains	96.0
Alr8	D1a <sup>d</sup>				d5my6b1	b.1.1.1: V set domains	95.7
Alr8	D1b <sup>e</sup>				d2esve1	b.1.1.1: V set domains	96.0
Alr9	D1				d5my6b1	b.1.1.1: V set domains	96.1
Alr11	D1				d5my6b1	b.1.1.1: V set domains	96.7
Alr12A	D1				d4n8pa1	b.1.1.1: V set domains	96.6
Alr12B	D1				d5my6b1	b.1.1.1: V set domains	96.9
Alr15	D1						
Alr16	D1				d5my6b1	b.1.1.1: V set domains	96.9
Alr17	D1				d5my6b1	b.1.1.1: V set domains	97.1
Alr18	D1				d5my6b1	b.1.1.1: V set domains	97.0
Alr19	D1				d5my6b1	b.1.1.1: V set domains	96.4
Alr21	D1	PF07686.17	V-set	0.0012	d5my6b1	b.1.1.1: V set domains	96.8
Alr23	D1				d5my6b1	b.1.1.1: V set domains	96.9
Alr27	D1						
Alr28	D1	PF07686.17	V-set	1.40E-04	d5o04f1	b.1.1.1: V set domains	95.3
Alr29	D1	PF07686.17	V-set	8.10E-05	d1yjdc1	b.1.1.1: V set domains	95.5
Alr30	D1	PF07686.17	V-set	0.0031	d5e56a	b.1.1.1: V set domains	95.0
Alr31	D1	PF17711.1	DUF5556	0.0097			
Alr33	D1						
Alr34	D1	PF07686.17	V-set	1.00E-04	d1c5db1	b.1.1.1: V set domains	88.8
Alr35	D1				d5o04f1	b.1.1.1: V set domains	93.9
Alr36	D1				d5my6b1	b.1.1.1: V set domains	93.4
Alr38	D1				d5my6b1	b.1.1.1: V set domains	69.5

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

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

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

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

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

Homologous proteins can evolve such that their primary sequences become highly divergent but their tertiary structures remain relatively unchanged (A.-S. Yang & Honig, 2000). Thus, structural alignments can often reveal distant homologies that cannot be detected by sequence-based methods. Therefore, we predicted the three-dimensional structure of each domain 1 sequence with AlphaFold (Jumper et al., 2021), as implemented in Colabfold (Mirdita et al., 2021). AlphaFold is a deep learning algorithm 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 AlphaFold 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 domain 1 sequences, the structural predictions had average (model-wide) pLDDT scores ranging from 80.6 to 97.4, with 22/29 models >90 (Table 5 and Figure 23). Thus, AlphaFold was able to confidently predict the backbones of all domain 1 folds.

**Table 5. Sequence homology and predicted structural homology of Domain 1.**

Protein <sup>a</sup>	AlphaFold <i>plDDT score</i> <sup>b</sup>	DALI Top Structural Alignment					TM-align <i>TM-score</i> <sup>f</sup>	Domain Type
		<i>PDB accession</i>	<i>Model</i>	<i>Z-score</i> <sup>c</sup>	<i>LALI</i> <sup>d</sup>	<i>RMSD</i> <sup>e</sup>		
Alr1	97.4	<a href="#">7kqy-E</a>	Antibody heavy chain	15.2	109	1.9	0.79	V-set
Alr2	90.5	<a href="#">3oai-A</a>	Myelin protein P0	15.1	108	2.1	0.82	V-set
Alr3	95.2	<a href="#">3udw-D</a>	Poliovirus receptor	14.8	107	1.9	0.73	V-set
Alr4	92.7	<a href="#">5imk-A</a>	V-set and Ig domain-containing protein 4	14.4	110	1.8	0.80	V-set
Alr6	97.0	<a href="#">6o3b-E</a>	Antibody heavy chain	15.7	103	1.5	0.80	V-set
Alr7	92.5	<a href="#">2ice-S</a>	V-set and Ig domain-containing protein 4	14.9	117	2.1	0.83	V-set
Alr8 a <sup>g</sup>	94.7	<a href="#">2ice-T</a>	V-set and Ig domain-containing protein 4	15.2	114	2.1	0.83	V-set
Alr8 b <sup>h</sup>	89.3	<a href="#">5imk-A</a>	V-set and Ig domain-containing protein 4	13.5	107	2.1	0.77	V-set
Alr9	96.9	<a href="#">6o3b-E</a>	Antibody heavy chain	14.7	100	1.5	0.78	V-set
Alr11	96.7	<a href="#">6o3b-E</a>	Antibody heavy chain	15.7	103	1.6	0.81	V-set
Alr12A	95.4	<a href="#">5imk-A</a>	V-set and Ig domain-containing protein 4	14.7	106	1.8	0.83	V-set
Alr12B	92.8	<a href="#">5imk-A</a>	V-set and Ig domain-containing protein 4	14.9	109	2	0.83	V-set
Alr15	85.6	<a href="#">6bj2-D</a>	TCR 589 alpha chain	14	108	2.1	0.75	V-set
Alr16	95.7	<a href="#">2ice-T</a>	V-set and Ig domain-containing protein 4	14.6	112	2.1	0.81	V-set
Alr17	95.5	<a href="#">2ice-S</a>	V-set and Ig domain-containing protein 4	15	114	2.1	0.80	V-set
Alr18	95.7	<a href="#">3qi9-D</a>	NKT TCR V beta 6 2A3-D	14.9	105	1.8	0.79	V-set
Alr19	97.3	<a href="#">1tvd-B</a>	T-cell Receptor, delta chain	15.3	109	2.2	0.78	V-set
Alr21	95.0	<a href="#">6j8g-C</a>	Sodium channel subunit beta-2	14.4	112	2	0.82	V-set
Alr23	95.9	<a href="#">2pnd-A</a>	V-set and Ig domain-containing protein 4	15.3	108	1.7	0.83	V-set
Alr27	84.2	<a href="#">2f53-D</a>	T-cell Receptor, alpha chain	14.1	107	2	0.76	V-set
Alr28	80.6	<a href="#">5m2w-B</a>	Llama nanobody nb8	14	108	2.1	0.78	V-set
Alr29	88.7	<a href="#">2ice-T</a>	V-set and Ig domain-containing protein 4	17.1	109	1.6	0.87	V-set
Alr30	92.8	<a href="#">3oai-A</a>	Myelin protein P0	15.8	108	2	0.85	V-set
Alr31	86.7	<a href="#">2ice-T</a>	V-set and Ig domain-containing protein 4	14.2	105	2	0.80	V-set
Alr33	84.3	<a href="#">6dle-B</a>	IgLON family member 5	12.6	91	1.6	0.68	Ig-like
Alr34	95.7	<a href="#">5imk-A</a>	V-set and Ig domain-containing protein 4	15	112	2.4	0.72	V-set
Alr35	93.5	<a href="#">1tvd-A</a>	T-cell Receptor, delta chain	16.7	114	2.2	0.81	V-set
Alr36	92.9	<a href="#">6fr6-B</a>	T-cell Receptor Beta Chain	15.3	107	1.9	0.77	V-set
Alr38	93.4	<a href="#">5iml-A</a>	V-set and Ig domain-containing protein 4	15.8	115	2.2	0.81	V-set

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

<sup>e</sup> RMSD < 2 are considered reasonable models

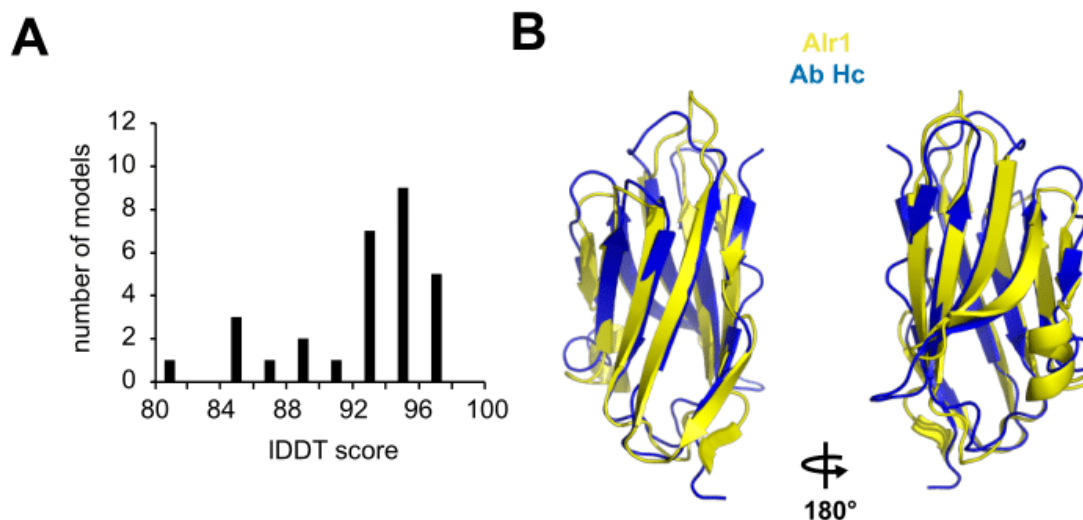
<sup>b</sup> predicted local-distance difference test; values >80 are considered good models <sup>f</sup> TM-scores > 0.5 indicate proteins with same general topology and are shaded green

<sup>c</sup> Z-score between 8-20 are considered probably homologous

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

<sup>d</sup> LALI = Number of equivalent residues considered in Z-score

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



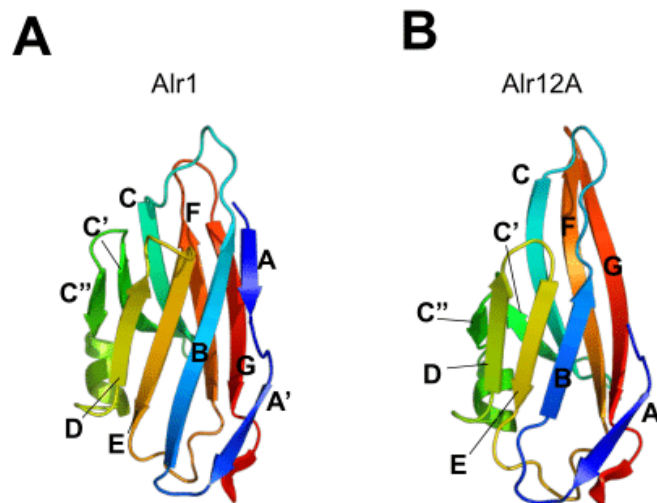
**Figure 23. Structural predictions of domain 1.**

A) Histogram of average pLDDT scores for AlphaFold structural models of domain 1. Models with an average pLDDT score of >70 are considered to be reasonable model while scores >90 indicate a that is highly likely to have a correct topology. B) Structural alignments of Alr1 domain 1 to a V-set Ig domain from the Human Antibody Heavy Chain (Ab Hc; PDB 7KQY chain E).

We then used Dali (Holm, 2020) to compare each domain 1 structure to the full Protein Data Bank (PDB; rcsb.org; Berman et al., 2000). 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 > 20 are definitely homologous, and Z-scores between 8-20 are probably homologous. For domain 1, the top hit for each model had a Z-score ranging from 12.6-16.7. In 28/29 cases, the top hit was to a V-set Ig domain, with the remaining model (Alr33) aligned to an Ig domain that, itself, could not be classified as a specific subtype by Interproscan (data not shown). To assess global structural similarity between each domain 1 model and its top hit, we aligned them with TMalign (Zhang & Skolnick, 2005). TMalign assigns structural alignments a TM-score ranging from 0 to 1, where 1 indicates a perfect match, <0.2 corresponds to unrelated proteins, and >0.5

indicates proteins that have the same general topology. All domain 1 models aligned to their top hit with TM-scores ranging from 0.68 to 0.87 (Table 5).

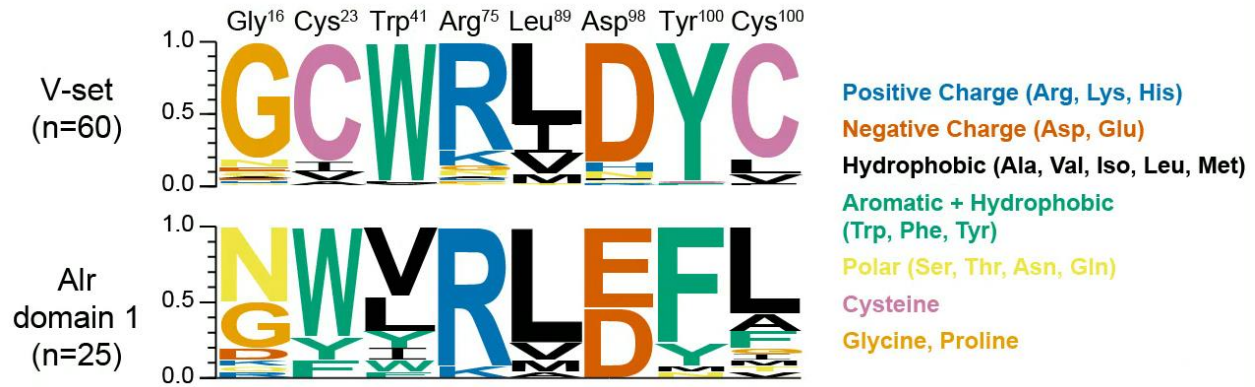
To further explore the similarity of domain 1 to canonical V-set Ig-like domains, I investigated whether they shared the same number and arrangement of  $\beta$ -strands. V-set domains have nine  $\beta$ -strands named A, B, C, C', C'', D, E, F, and G according to their position in the primary amino acid sequence. In V-set domains, strand A is usually split into A and A'. Strand C'' is only found in V-set domains. The nine  $\beta$ -strands are arranged as a Greek key and form a  $\beta$ -sandwich, such that one  $\beta$ -sheet is formed by the A,B,E, and D strands, and the other is formed by the A', G, F, C, C', and C'' strands. This arrangement is often referred to as the V-frame (Harpaz & Chothia, 1994). I used PyMOL to visually inspect each structural model and found that the protein backbone traced the path of a V-frame Greek key (e.g. Figure 24A,B). This was not surprising because the structural alignments had already indicated a good fit with V-set domains. I then used STRIDE (Frishman & Argos, 1995) to predict the secondary structure of each model, and assigned letters to the  $\beta$ -strands according to their order in the primary sequence and their position in the fold. I found that 25/29 models were predicted to have the nine V-frame  $\beta$ -strands. Four models were predicted to have eight strands, with the missing strand being either the A or A' position (e.g., see Figure 24B). Notably, all models had the V-set specific C'' strand (e.g., Figure 24A). Thus, the similarity of domain 1 to V-set Ig-like domains extends to their secondary structure.



**Figure 24. Predicted topology of  $\beta$ -strands of domain 1.**

$\beta$ -strands labeled in A) Alr1 domain 1, and B) Alr12A domain 1.

The fact that domain 1 sequences were predicted to fold like a V-set domain led us to question why HMMER did not identify them as such. 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 (Harpaz & Chothia, 1994; Litman et al., 2001; Cannon et al., 2002). According to the nomenclature of Cannon et al (2002), these residues are Gly<sup>16</sup>, Cys<sup>23</sup>, Trp<sup>41</sup>, Arg<sup>75</sup>, Leu<sup>89</sup> (or other hydrophobic residue), Asp<sup>98</sup>, Tyr<sup>102</sup>, and Cys<sup>104</sup>. To determine whether these residues are conserved in the 25 domain 1 sequences most similar to V-set domains, I generated a multiple sequence alignment between them and 60 canonical V-set sequences from the Pfam V-set sequence profile (pf07686). Then I identified the domain 1 residues that corresponded to the eight V-set residues. My findings are summarized as a sequence logo in Figure 25.

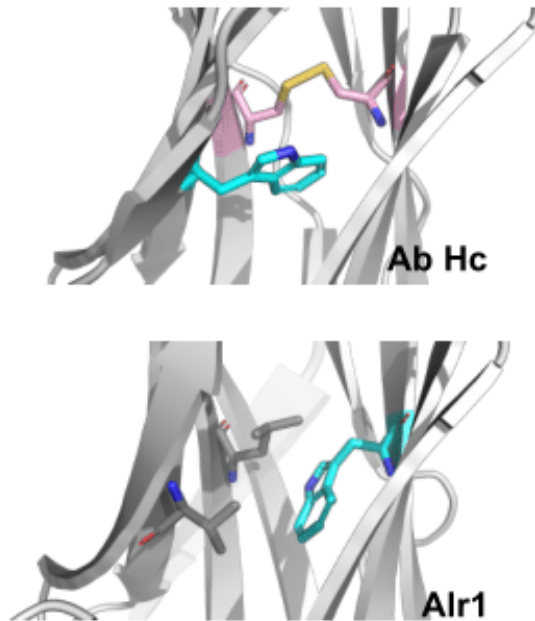


**Figure 25. Sequence logo of amino acid frequencies in canonical V-set and *Alr* domain 1.**

Sequence logo comparison at the eight conserved positions in V-set Ig domains (top) and *Alr* domain 1 sequences (bottom).

In V-set domains, Cys<sup>23</sup>, Trp<sup>41</sup>, and Cys<sup>104</sup> form a nearly invariant structural motif called the ‘pin’ (Lesk & Chothia, 1982). The cysteines form a disulfide linkage between  $\beta$ -strands B and F, while the aromatic side chain of the tryptophan packs against the bond to stabilize the hydrophobic core of the  $\beta$ -sandwich (e.g. Figure 26, top). Although a few canonical Ig-like domains lack either the cysteines or the tryptophan, V-set domains that lack all three are unheard of. In contrast, all *Alr* domain 1 sequences lacked Cys residues at positions 23 and 104, and only 2/29 had a Trp at position 41. Instead, Cys<sup>23</sup> was replaced by hydrophobic amino acids with bulky, aromatic side chains (Trp, Phe, or Tyr). Cys<sup>104</sup> and Trp<sup>41</sup> were replaced by hydrophobic amino acids. Figure 26 shows an example from domain 1 of *Alr1*. 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  $\beta$ -sandwich.

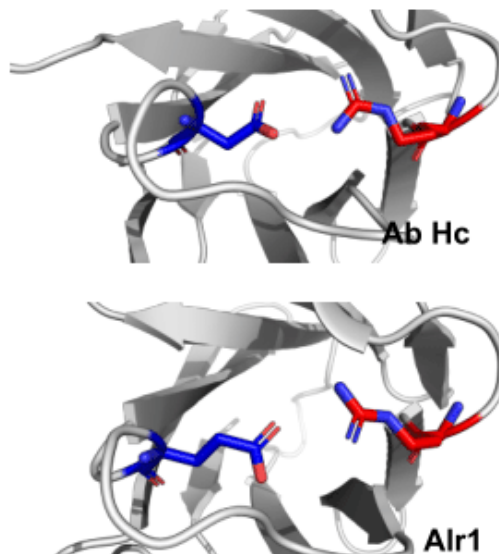




**Figure 26. CWC “pin” motif comparison.**

Arrangement of amino acids forming the “pin” motif in the crystal structure of human antibody heavy-chain (Ab Hc) (top) and the predicted structure of Alr1 domain 1 (bottom).

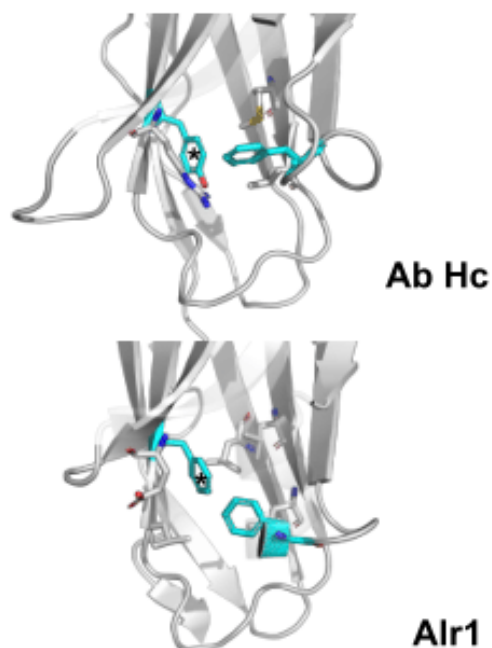
The fourth and fifth V-set residues, Arg<sup>75</sup> and Asp<sup>98</sup>, form a salt bridge between the CD and EF loops. The salt bridge is thought to stabilize the “bottom” of the domain and is found only in the V-set and I-set immunoglobulin domains (Harpaz & Chothia, 1994; Cannon et al., 2002). I found the salt bridge in all but three (26/29) Alr domain 1 models, although the negatively charged Asp was often replaced with a similarly charged Glu (Figure 25, Figure 27). Thus, the salt bridge, a hallmark of V-set and I-set domains, is also present in domain 1.



**Figure 27. Salt bridge comparison.**

Presence of a salt bridge connecting the CD and EF loops in the structure of human antibody heavy-chain (Ab Hc) (top) and the predicted structure of Alr1 domain 1 (bottom).

The sixth canonical residue, Tyr<sup>102</sup>, 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). Tyr<sup>102</sup> stabilizes the  $\beta$ -sandwich via hydrophobic interactions between its aromatic group and other side chains in the core of the fold (Figure 28). Its hydroxyl group also forms hydrogen bonds to stabilize the EF loop, one of three topologically important loops that cross the  $\beta$ -sandwich (Hamill et al., 2000). While Tyr<sup>102</sup> is highly conserved in V-set Ig-like domains (97% in my seed alignment), it was found in only 7/29 Alr domains (Figure 25). Instead, 20/29 had Phe, with its aromatic ring occupying the same location as that of Tyr<sup>102</sup> (Figure 28). Mutational studies have shown that a Tyr to Phe 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 do not rule out the possibility that domain 1 folds like a V-set Ig domain.

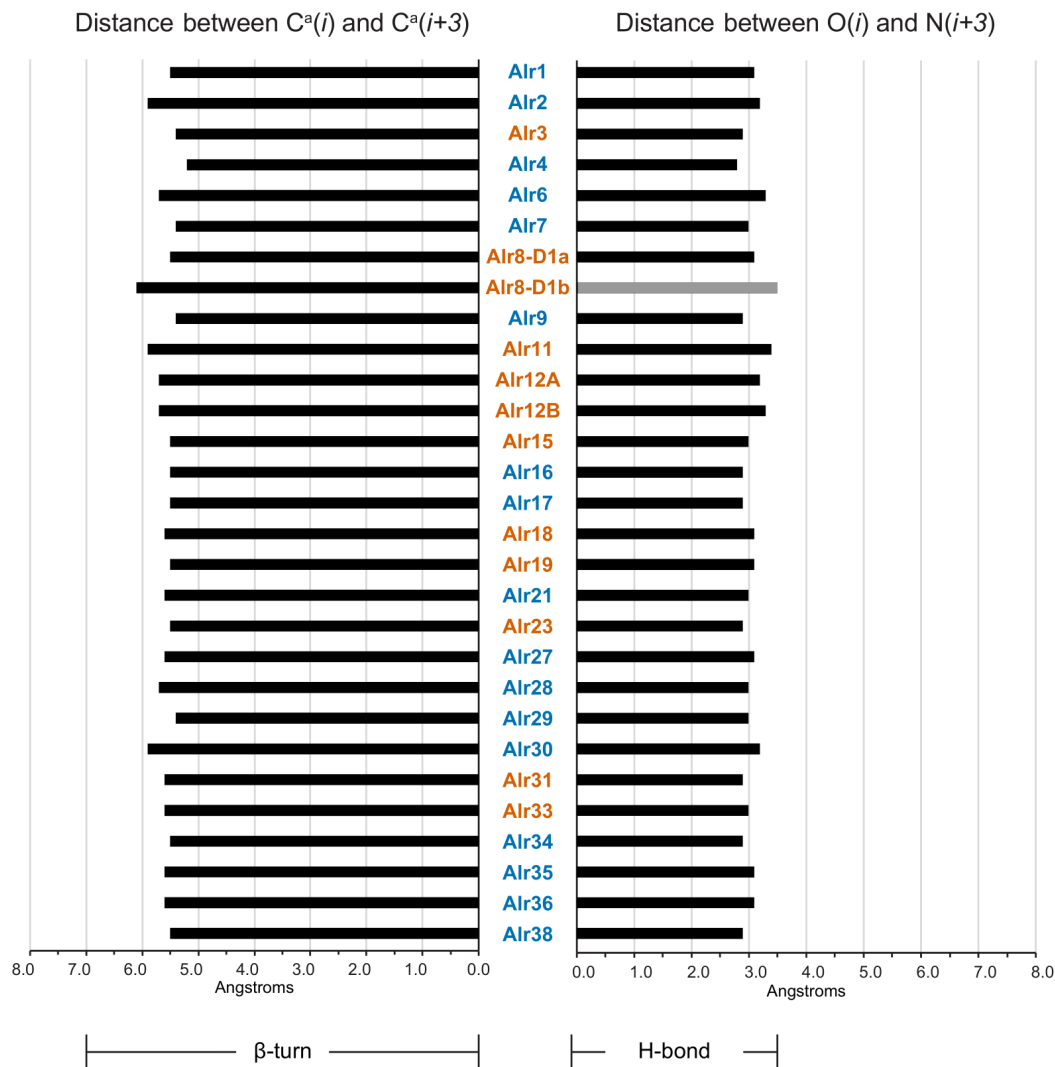


**Figure 28. Hydrophobic core comparison.**

Orientation of the tyrosine corner in the crystal structure of human antibody heavy-chain (Ab Hc) (top), and the corresponding region in Alr1 domain 1 (bottom). The conserved Tyr (and the Phe that replaces it in Alr1) are indicated with an asterisk. Nearby amino acids with inward facing hydrophobic residues are also shown.

The seventh canonical residue is Gly<sup>16</sup>, which is part of a  $\beta$ -turn between strands A' and B. A  $\beta$ -turn is a series of four residues that reverses 180° on itself such that the distance between C <sup>$\alpha$</sup> ( $i$ ) and C <sup>$\alpha$</sup> ( $i+3$ ) is less than 7 Å (Chou, 2000).  $\beta$ -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). Glycine is common in the second and third positions of  $\beta$ -turns; in V-set domains, Gly<sup>16</sup> is located at position  $i + 2$ . Within the models of domain 1, all 29 were predicted to have a  $\beta$ -turn (Figure 29). Twenty-eight of these featured a hydrogen bond, defined by the criterion that O( $i$ ) is < 3.5 Å from N( $i+3$ ) (Richardson, 1981). Position  $i + 2$  was Gly in eleven sequences, Asn in thirteen sequences, Asp in two cases, and Arg, Lys, or Gln in one

sequence each. Thus, like V-set domains, domain 1 is predicted to have a  $\beta$ -turn between strands A' and B, but in most cases it does not involve a glycine.



**Figure 29. Predicted  $\beta$ -turn and hydrogen bonds in Alr D1 structures.**

Bona fide *Alr* genes are color coded in blue and the putative *Alr* genes are color coded in orange.

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 domain 1 sequences.

In summary, data based on comparisons of the primary sequence and predicted secondary and tertiary structures of domain 1 are consistent with most of them being homologous to V-set Ig domains.

### **3.4.2 Domains 2 and 3 are IgSF domains most similar to the I-set family**

Domains 2 and 3 of Alr1 and Alr2 were originally described as similar to I-set Ig-like domains (Nicotra et al., 2009; Rosa et al., 2010). We expanded this analysis to the 29 domain 2 and two domain 3 sequences encoded by bona fide and putative Alr genes. First, I used HMMER to compare all 31 sequences to the Pfam database. At an E-value cutoff of  $<0.01$ , 14 were identified as I-set Ig-like domains (pf07679) and another four as generic Ig-like domains (pf13927) (Table 6). Then, I used HHpred to compare each sequence to the SCOPe database. Twenty-three of thirty-one domains had a  $>95\%$  probability of being homologous to the I-set family of Ig domains (Table 6).

**Table 6. Sequence homology Domain 2 and 3.**

Protein <sup>a</sup>	Domain	HMMER search of pfam			HHpred search of SCOPe		
		Accession	description	e-value <sup>b</sup>	Accession	SCOPe family	probability <sup>c</sup>
Alr1	D2	PF07679.16	I-set	3.30E-05	d1biha3	b.1.1.4: I-set domains	97.6
Alr2	D2	PF13927.6	Ig_3	0.0002	d1biha3	b.1.1.4: I-set domains	99.0
Alr2	D3	PF07679.16	I-set	0.0023	d1biha3	b.1.1.4: I-set domains	96.7
Alr3	D2				d1biha3	b.1.1.4: I-set domains	96.6
Alr4	D2				d1biha3	b.1.1.4: I-set domains	98.7
Alr6	D2	PF07679.16	I-set	1.70E-05	d1biha3	b.1.1.4: I-set domains	97.0
Alr7	D2	PF07679.16	I-set	8.80E-05	d1x44a1	b.1.1.4: I-set domains	99.3
Alr8	D2a <sup>d</sup>	PF07679.16	I-set	0.0024	d1biha3	b.1.1.4: I-set domains	97.4
Alr8	D2b <sup>e</sup>	PF07679.16	I-set	2.20E-07	d1biha3	b.1.1.4: I-set domains	98.1
Alr9	D2				d1biha3	b.1.1.4: I-set domains	97.4
Alr11	D2	PF07679.16	I-set	0.0034	d1biha3	b.1.1.4: I-set domains	97.1
Alr12A	D2	PF13927.6	Ig_3	0.0028	d1biha3	b.1.1.4: I-set domains	97.0
Alr12B	D2	PF13927.6	Ig_3	0.0024	d1biha3	b.1.1.4: I-set domains	97.2
Alr15	D2				d1biha3	b.1.1.4: I-set domains	84.9
Alr16	D2	PF07679.16	I-set	0.0021	d1biha3	b.1.1.4: I-set domains	97.4
Alr18	D2	PF07679.16	I-set	3.50E-06	d1biha3	b.1.1.4: I-set domains	97.4
Alr19	D2	PF07679.16	I-set	2.50E-06	d1biha3	b.1.1.4: I-set domains	97.4
Alr21	D2	PF07679.16	I-set	8.10E-05	d1biha3	b.1.1.4: I-set domains	97.6
Alr23	D2	PF07679.16	I-set	0.0005	d1biha3	b.1.1.4: I-set domains	97.5
Alr27	D2						
Alr28	D2	PF07679.16	I-set	0.0076	d1vcaa2	b.1.1.4: I-set domains	97.4
Alr29	D2				d1ncua1	b.1.1.4: I-set domains	85.2
Alr30	D2	PF13927.6	Ig_3	0.00027	d1biha3	b.1.1.4: I-set domains	97.8
Alr30	D3				d1biha3	b.1.1.4: I-set domains	96.9
Alr31	D2				d1ncua1	b.1.1.4: I-set domains	87.2
Alr33	D2				d1iray3	b.1.1.4: I-set domains	54.4
Alr34	D2				d1biha3	b.1.1.4: I-set domains	97.1
Alr35	D2	PF07679.16	I-set	0.0066	d1koa1	b.1.1.4: I-set domains	89.8
Alr36	D2				d1biha3	b.1.1.4: I-set domains	97.4
Alr37	D2				d1biha3	b.1.1.4: I-set domains	78.7
Alr38	D2				d1iray3	b.1.1.4: I-set domains	54.7

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

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

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

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

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

I then predicted the structure of each domain 2 and domain 3 with AlphaFold. All models had an average pLDDT > 80, with 19/31 having an average pLDDT >90 (Table 7 and Figure 30A). To determine whether these models had the same topology as an I-set Ig fold, we predicted their secondary structure using STRIDE then visualized them in PyMOL. The topology of an I-set Ig domain is similar to a V-set domain except that it lacks a C'' strand, and the C' strand is typically shorter (Harpaz & Chothia, 1994). We found that 21 of our models had an I-set topology (Figure 30B). Two additional models (Alr12B and Alr33) also had I-set topologies except that a beta-strand was not predicted in the location of either A or A'. Eight other models had an I-set-like topology but were missing the C' strand. Thus, although most domain 2 and domain 3 sequences are predicted fold like an I-set domain, several appear to lack the C' strand.

**Table 7. Sequence homology and predicted structural homology of Domain 2 and Domain 3.**

Protein <sup>a</sup>	AlphaFold pLDDT score <sup>b</sup>	DALI Top Structural Alignment					TM-align TM-score <sup>f</sup>	Domain Type
		PDB accession	Model	Z-score <sup>c</sup>	LALI <sup>d</sup>	RMSD <sup>e</sup>		
Alr1 D2	95.1	<a href="#">2rjm-A</a>	Ig domains from Titin	12.9	89	1.7	0.78	I-set
Alr2 D2	93.7	<a href="#">1u2h-A</a>	Aortic preferentially expressed protein 1	12.4	88	1.7	0.81	I-set
Alr2 D3	90.9	<a href="#">6efy-A</a>	Dpr-interacting protein alpha, isoform A	12.9	90	1.5	0.84	I-set
Alr3 D2	88.2	<a href="#">2rik-A</a>	I-band fragment from Titin	12.7	89	1.8	0.75	I-set
Alr4 D2	94.4	<a href="#">2j8h-A</a>	Ig repeat from Titin	13.2	87	1.4	0.80	I-set
Alr6 D2	93.9	<a href="#">2rjm-A</a>	Ig domains from Titin	12.2	87	1.7	0.76	I-set
Alr7 D2	92.1	<a href="#">2rjm-A</a>	Ig domains from Titin	13	88	1.5	0.79	I-set
Alr8 a D2	88.0	<a href="#">2rjm-A</a>	Ig domains from Titin	13.4	91	1.8	0.79	I-set
Alr8 b D2	92.5	<a href="#">2rjm-A</a>	Ig domains from Titin	13.6	89	1.4	0.80	I-set
Alr9 D2	81.1	<a href="#">4pgz-A</a>	Mast/stem cell growth factor receptor Kit	10.5	96	2.6	0.71	Ig-like
Alr11 D2	88.2	<a href="#">2ill-A</a>	Substructure of Titin	12.1	87	1.6	0.77	I-set
Alr12A D2	88.4	<a href="#">3puc-A</a>	Titin domain M7	12.9	87	1.6	0.78	I-set
Alr12B D2	85.7	<a href="#">4of8-B</a>	Irregular chiasm C-roughest protein	12	97	2.1	0.78	C2-set
Alr15 D2	92.6	<a href="#">4of8-B</a>	Irregular chiasm C-roughest protein	10.8	93	2.1	0.79	C2-set
Alr16 D2	85.9	<a href="#">4uow-5</a>	Titin M10-Obscurin Ig domain 1 complex	11.5	85	2.1	0.71	I-set
Alr18 D2	89.4	<a href="#">2rjm-A</a>	Ig domains from Titin	13.1	90	1.7	0.79	I-set
Alr19 D2	92.6	<a href="#">2rjm-A</a>	Ig domains from Titin	12.1	87	1.8	0.76	I-set
Alr21 D2	87.5	<a href="#">6efy-A</a>	Dpr-interacting protein alpha, isoform A	12.9	93	2	0.81	I-set
Alr23 D2	92.2	<a href="#">4pgz-B</a>	Mast/stem cell growth factor receptor Kit	11.6	94	2.3	0.74	Ig-like
Alr27 D2	92.4	<a href="#">3sbw-C</a>	Programmed cell death 1 ligand 1	10.7	88	2.3	0.74	C2-set
Alr28 D2	91.8	<a href="#">4pgz-B</a>	Mast/stem cell growth factor receptor Kit	13	93	1.9	0.83	Ig-like
Alr29 D2	86.9	<a href="#">4of8-B</a>	Irregular chiasm C-roughest protein	10.6	92	2.2	0.77	C2-set
Alr30 D2	90.4	<a href="#">1u2h-A</a>	Aortic preferentially expressed protein 1	12.9	90	1.6	0.83	I-set
Alr30 D3	92.1	<a href="#">2fdb-P</a>	Fibroblast growth factor receptor 2	12.3	90	1.8	0.80	I-set
Alr31 D2	84.0	<a href="#">3dmk-C</a>	DSCAM 1.30.30, N-terminal Ig domains	11.4	93	2.4	0.74	I-set
Alr33 D2	89.0	<a href="#">6pv9-A</a>	Programmed cell death 1 ligand 1	10.4	85	2.2	0.71	C2-set
Alr34 D2	93.4	<a href="#">2j8h-A</a>	Ig repeat from Titin	12.8	87	1.5	0.81	I-set
Alr35 D2	95.5	<a href="#">3dmk-C</a>	DSCAM 1.30.30, N-terminal Ig domains	12.4	92	2.3	0.75	I-set
Alr36 D2	90.0	<a href="#">2rik-A</a>	I-band fragment from Titin	13.4	90	1.7	0.79	I-set
Alr37 D2	92.6	<a href="#">4uow-R</a>	Titin M10-Obscurin Ig domain 1 complex	10.9	84	2.2	0.74	I-set
Alr38 D2	91.7	<a href="#">2rik-A</a>	I-band fragment from Titin	12.5	87	1.7	0.79	I-set

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

<sup>b</sup> predicted local-distance difference test; values >80 are considered good models

<sup>c</sup> Z-score between 8-20 are considered probably homologous

<sup>d</sup> LALI = Number of equivalent residues considered in Z-score

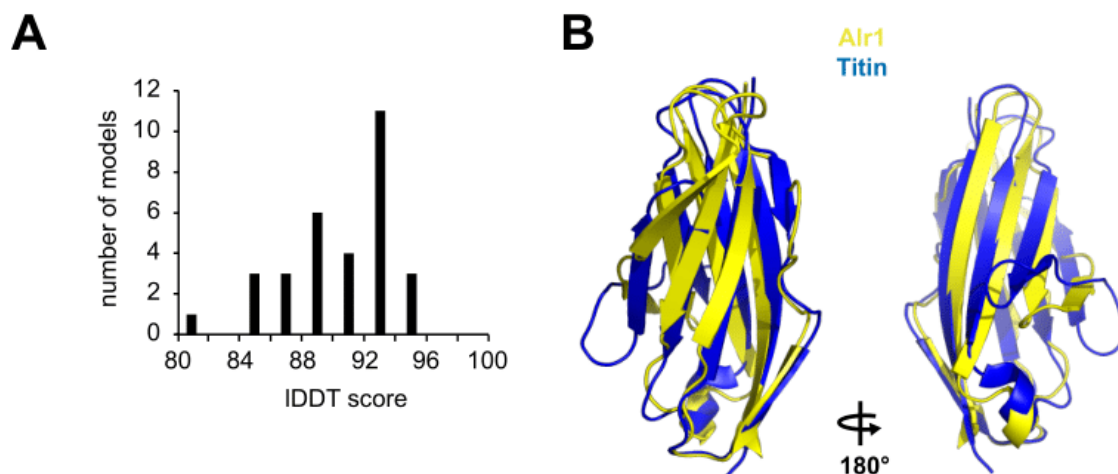
<sup>e</sup> RMSD < 2 are considered reasonable models

<sup>f</sup> TM-scores > 0.5 indicate proteins with same general topology and are shaded green

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

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

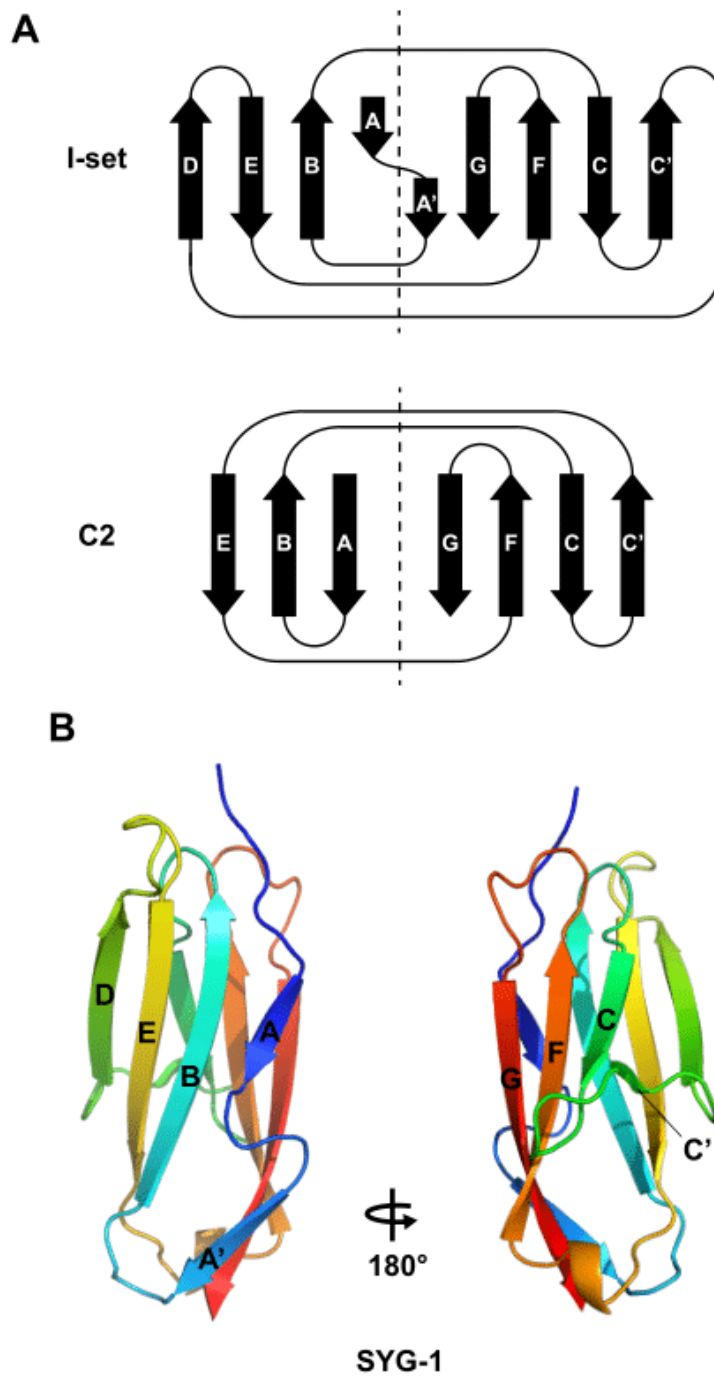




**Figure 30. Structural predictions of domains 2 and 3.**

A) Histogram of pLDDT scores for AlphaFold structural models of domains 2 and 3. B) Structural alignment of Alr1 domain 2 to an I-set Ig domain from Rabbit Titin (PDB 2RJM, central domain).

When we searched the PDB for proteins structurally similar to domains 2 and 3, we found that most of the top hits produced by Dali were I-set domains (Table 7 and Figure 31). However, several top hits were labeled as C2-set domains according to Pfam (Table 7). A C2-set Ig domain has seven beta-strands and lacks the D strand (van Sorge et al., 2021) (Figure 31A). Thus, one beta-sheet is formed by strands A, B, and E, and the other by strands G, F, C, and C'. Surprisingly, when we inspected the structures of these C2 domains we found their topology was actually that of an I-set domain. For example, the top hit for Alr12B domain 2 was to an Ig domain from SYG-1, a cell adhesion molecule in *Drosophila* (PDB 4of8, chain B). It has a D strand, and its 9 beta-strands follow the I-set topology (Figure 31B). We found similar results for all of the “C2-set” hits produced by Dali. For all models, the top hits in Dali had Z-scores ranging from 10.4 to 13.4, indicating probable homology. Thus, domains 2 and 3 are most similar to Ig domains with an I-set topology.

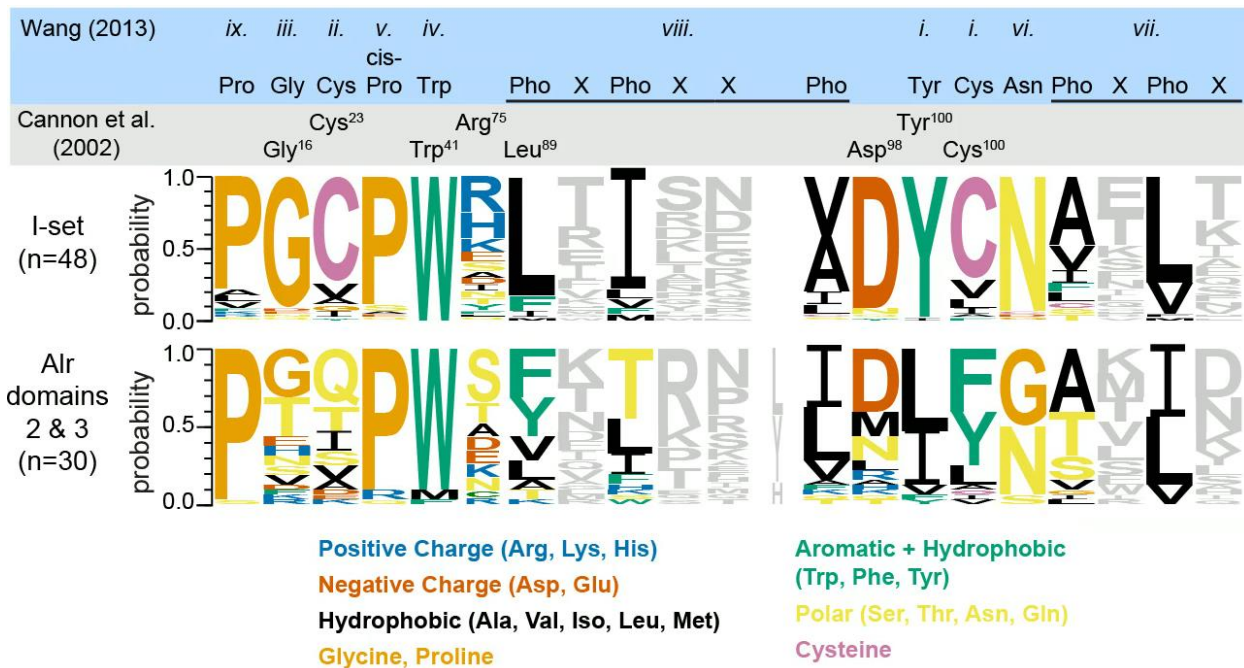


**Figure 31. Topology of  $\beta$ -strands of I-set and C2 domains.**

A) Greek key topology of  $\beta$ -strands in I-set domains. B) Predicted topology of  $\beta$ -strands in SYG-1 (PDB 4of8, chain

B) shows that it is an I-set domain and not a C2 domain.

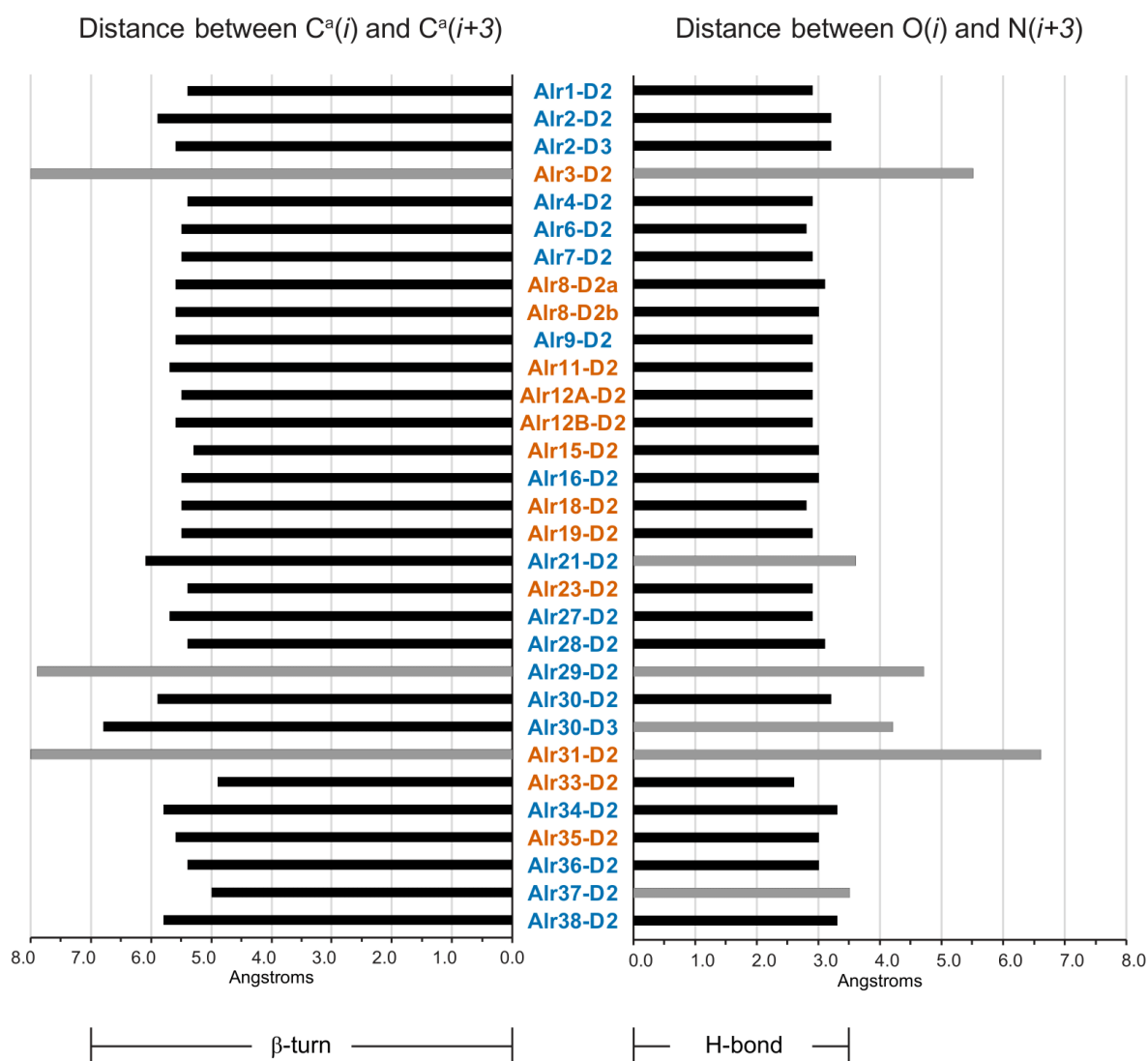
Next, I investigated whether domains 2 and 3 had any of the conserved sequence motifs found in I-set domains. To do so, we aligned the domains to 48 canonical I-set domains from the Pfam I-set sequence profile (pf07679). We then searched for the sequence motifs common to V-frame Ig-like domains (Harpaz & Chothia, 1994; Cannon et al., 2002). With respect to the pin motif (C-W-C), the Alr domains had the central tryptophan (or a bulky hydrophobic residue) but lacked the paired cysteines. One cysteine was replaced by a hydrophobic residue in the Alr domains, while the second was replaced by residues bearing no consistent physicochemical property (Figure 32). The Alr domains also lacked the salt bridge and tyrosine corner (Figure 32). The  $\beta$  turn between  $\beta$ -strands A' and B was present in all but three structural models. The last of the eight conserved residues, a hydrophobic residue (typically leucine) was present, although in many cases a tyrosine was present in that eighth position (Figure 32).



**Figure 32. Sequence logo of amino acid frequencies in canonical I-set and Alr domain 2 and 3.**

Sequence logo comparing frequency of amino acids at conserved positions in I-set Ig domains (top) and Alr domains 2 and 3 (bottom). Eight conserved V-frame positions are highlighted with a gray background, while additional I-set motifs are indicated with a blue background.

More recently, the sequence signature of I-set domains was defined via nine sequence motifs, denoted *i* through *ix* (J. H. Wang, 2013). 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*) (Figure 32 and Figure 34). However, the remaining five motifs had not been discussed in previous analyses (Harpaz & Chothia, 1994; Litman et al., 2001; Cannon et al., 2002) so I searched for these motifs in domains 2 and 3.



**Figure 33. Predicted  $\beta$ -turn and hydrogen bonds in Alr D2 structures.**

Bona fide Alr genes are color coded in blue and the putative Alr genes are color coded in orange.

Motif  $v$  is a conserved proline in the BC loop, typically located six positions upstream of the conserved tryptophan. Motif  $vi$  is a conserved asparagine in the FG loop. These two residues form a hydrogen bond that stabilizes the BC and FG loops in a closed position. In domains 2 and 3, I found the conserved proline residue in 28/31 sequences but the asparagine was only present in

13/31 sequences (Figure 32). Furthermore, I found hydrogen bonds between the BC and FG loops in only 9/31 models. Thus, motif v is present in domains 2 and 3, but motif vi and the structural motif that it forms with motif v do not appear to be a common feature of these domains.

Motif *vii* is a Pho-X-Pho-X pattern of amino acids (where Pho represents a hydrophobic residue), located approximately 10-12 residues downstream of motif vi. It is found in  $\beta$ -strand G and denotes the C-terminal end of an I-set (and also V-set) domain. This pattern was found in 30/31 of domain 2 and 3 sequences (Figure 32).

Motif *viii* is also a set of hydrophobic and hydrophilic residues, Pho-X-Pho-X-X-Pho, located at the bottom of  $\beta$ -strand E. The last two hydrophobic residues typically make contact with the tyrosine in the tyrosine corner, while the two consecutive hydrophilic residues between them form a  $\beta$  bulge. In the alignments, 26/30 domain 2 and 3 sequences had this motif, although the first and second hydrophobic residues often had polar or aromatic side chains (Figure 32). As noted above, the tyrosine corner is not present in domains 2 and 3. However, a  $\beta$  bulge was predicted to occur at the end of the E strand in 28/31 structural models. Motif *viii* is therefore present in most domains, but the structural consequences of this motif are likely to differ from traditional I-set domains.

Motif *viii* is a proline ~23-26 residues upstream of the B-strand cysteine. This proline defines the beginning of an I-set domain and, in domains 2-3, this proline was found in 30/31 sequences (Figure 32).

Taken together, the data from the multiple sequence alignments and structural predictions indicate that domains 2 and 3 likely adopt an Ig fold with an arrangement of  $\beta$ -strands similar to that found in I-set domains. Yet, like domain 1, several of the sequence signatures and structural motifs traditionally used to identify I-set domains are missing in the *Hydractinia* sequences.

### **3.4.3 Part of the ECS adopts an immunoglobulin-like fold**

The ECS was originally described as the region between the extracellular Ig-like domains and the transmembrane helix of Alr1 and Alr2 (Nicotra et al., 2009; Rosa et al., 2010). This region was not found to be similar to other protein domains in HMMER or BLASTP searches. Here, I expanded the analysis to include the 29 ECS regions encoded by bona fide and putative Alr genes. HMMER searches against Pfam only returned one hit to a domain of unknown function. AlphaFold produced models of the entire ECS that aligned well to fibronectin type III (FN3) domains. To better define this potential immunoglobulin-like fold, 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 that the C-terminal portion did not. Using HHpred to compare each sequence to the SCOPe database, we found that 22/29 sequences only had 56-89% probability of homology with the FN3 family of Ig domains (Table 8). These results suggested that the N-terminal portion of the ECS may have a FN3-like fold.

**Table 8. Sequence homology of the ECS fold.**

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

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

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

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

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

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



As structure predictions are often more reliable when performed on single domains, we removed the excess C-terminal portion of the ECS sequences and repeated our HMMER and AlphaFold runs on the trimmed ECS. At an E-value cutoff of  $<0.01$ , HMMER only detected similarity between the Alr33 ECS and a domain of unknown function. However, AlphaFold confidently modeled all ECS sequences (Table 9 and Figure 34). Of these, 27/29 aligned significantly to FN3 domains with TM scores ranging from 0.70 to 0.88 (Table 9). The remaining two models showed significant alignments to FN3 superfamily immunoglobulin folds (Table 9). We included all 29 predicted models in subsequent analyses.

**Table 9. Sequence homology and predicted structural homology of ECS (trimmed).**

Protein <sup>a</sup>	AlphaFold <i>plDDT</i> score <sup>b</sup>	DALI Top Structural Alignment					TM-align <i>TM-score</i> <sup>f</sup>	Domain Type
		<i>PDB</i> accession	<i>Model</i>	<i>Z-score</i> <sup>c</sup>	<i>LALI</i> <sup>d</sup>	<i>RMSD</i> <sup>e</sup>		
Alr1	90.0	<a href="#">6h41-A</a>	Interleukin-5 receptor subunit alpha	11.4	82	1.8	0.82	FnIII SF
Alr2	95.6	<a href="#">5fn8-A</a>	CD45 d3-d4 <sup>i</sup>	12.4	85	1.9	0.79	FnIII
Alr3	94.1	<a href="#">5fn6-A</a>	CD45 d3-d4 <sup>i</sup>	11.5	83	1.9	0.78	FnIII
Alr4	95.4	<a href="#">7e9j-B</a>	POMGNT2 <sup>j</sup>	10.8	81	1.9	0.81	FnIII
Alr6	94.0	<a href="#">7e9k-D</a>	POMGNT2 <sup>j</sup>	11.2	80	1.8	0.84	FnIII
Alr7	90.0	<a href="#">5fn6-A</a>	CD45 d3-d4 <sup>i</sup>	10.8	80	1.9	0.81	FnIII
Alr8 <sup>g</sup>	91.1	<a href="#">7e9j-B</a>	POMGNT2 <sup>j</sup>	10.4	83	2.1	0.81	FnIII
Alr8 <sup>h</sup>	92.1	<a href="#">5fn6-A</a>	CD45 d3-d4 <sup>i</sup>	10.8	79	1.7	0.81	FnIII
Alr9	94.7	<a href="#">5fn8-A</a>	CD45 d3-d4 <sup>i</sup>	11.3	79	1.9	0.80	FnIII
Alr11	95.1	<a href="#">5fn8-A</a>	CD45 d3-d4 <sup>i</sup>	11.3	79	1.3	0.80	FnIII
Alr12A	95.3	<a href="#">5fn8-A</a>	CD45 d3-d4 <sup>i</sup>	10.8	79	1.9	0.79	FnIII
Alr12B	94.2	<a href="#">5x83-B</a>	Netrin receptor DCC	10.7	79	1.8	0.79	FnIII
Alr15	93.9	<a href="#">2gee-A</a>	Human Type III Fibronectin Extradomain B	11.6	80	2	0.80	FnIII
Alr16	93.7	<a href="#">5fn6-A</a>	CD45 d3-d4 <sup>i</sup>	11	80	1.7	0.82	FnIII
Alr18	93.1	<a href="#">6h41-A</a>	Interleukin-5 receptor subunit alpha	11.6	80	1.7	0.83	FnIII SF
Alr19	94.5	<a href="#">5fn6-A</a>	CD45 d3-d4 <sup>i</sup>	11.3	80	1.7	0.82	FnIII
Alr21	93.9	<a href="#">5fn6-A</a>	CD45 d3-d4 <sup>i</sup>	10.9	80	1.7	0.81	FnIII
Alr23	84.7	<a href="#">6xfi-A</a>	POMGNT2 <sup>j</sup>	10.7	84	2.2	0.83	FnIII
Alr27	94.3	<a href="#">2gee-A</a>	Human Type III Fibronectin Extradomain B	12.4	80	1.8	0.82	FnIII
Alr28	92.4	<a href="#">3t1w-A</a>	Four-domain fragment Fn7B89	12.4	85	1.6	0.81	FnIII
Alr30.1	88.2	<a href="#">3t1w-A</a>	Four-domain fragment Fn7B89	12.5	83	1.5	0.82	FnIII
Alr30.3	85.9	<a href="#">6moj-B</a>	Erythropoietin receptor	8.2	78	2.4	0.70	FnIII
Alr31	92.4	<a href="#">5fn8-B</a>	CD45 d3-d4 <sup>i</sup>	11.4	81	2	0.78	FnIII
Alr33	89.6	<a href="#">3t1w-A</a>	Four-domain fragment Fn7B89	11.6	87	2.1	0.79	FnIII
Alr34	92.3	<a href="#">5n48-D</a>	Fibronectin	12.7	88	1.8	0.78	FnIII
Alr35	96.5	<a href="#">5n48-D</a>	Fibronectin	11.5	83	2	0.79	FnIII
Alr36	93.1	<a href="#">5fn8-B</a>	CD45 d3-d4 <sup>i</sup>	10.9	80	1.9	0.80	FnIII
Alr37	92.0	<a href="#">5n48-D</a>	Fibronectin	11.9	86	1.9	0.79	FnIII
Alr38	92.2	<a href="#">5n48-D</a>	Fibronectin	12.5	87	2	0.79	FnIII

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

<sup>b</sup> predicted local-distance difference test; values >80 are considered good models

<sup>c</sup> Z-score between 8-20 are considered probably homologous

<sup>d</sup> LALI = Number of equivalent residues considered in Z-score

<sup>e</sup> RMSD < 2 are considered reasonable models

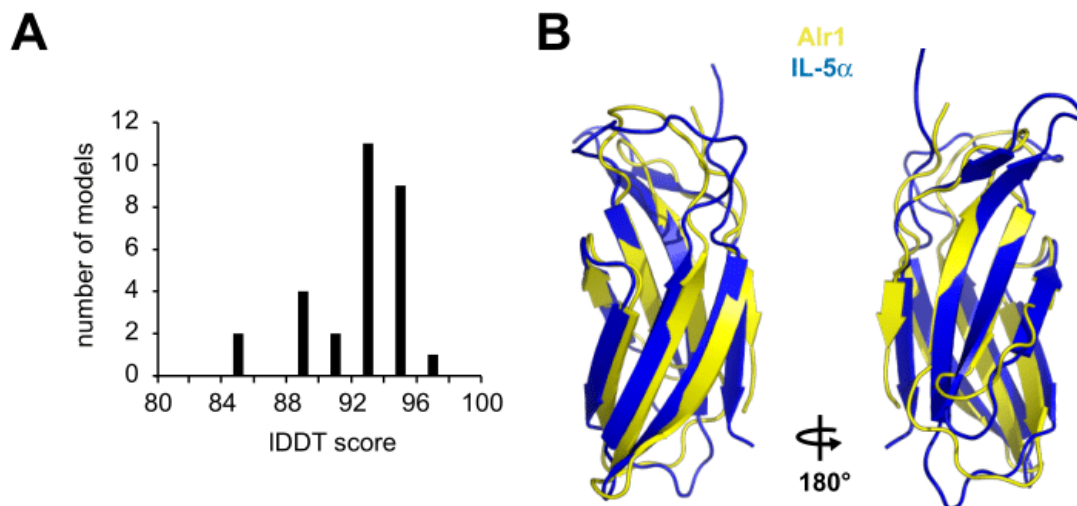
<sup>f</sup> TM-scores > 0.5 indicate proteins with same general topology and are shaded green

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

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

<sup>i</sup> Receptor-type tyrosine-protein phosphatase C

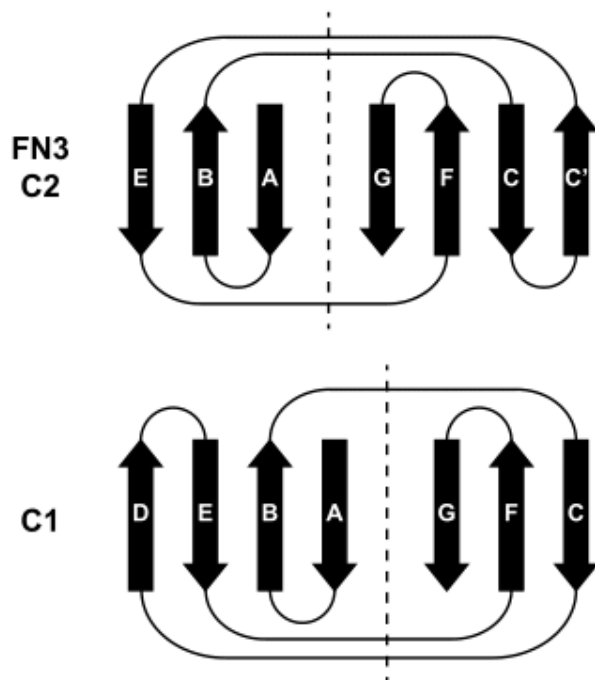
<sup>j</sup> Protein O-linked-mannose  $\beta$ -1,4-N-acetylglucosaminyltransferase 2



**Figure 34. Structural predictions of the ECS.**

A) Histogram of pLDDT scores for AlphaFold structural models of the region encoding an immunoglobulin-like fold. B) Structural alignment of the Alr1 ECS fold to human Interleukin-5 receptor subunit alpha (PDB 6H41, chain A).

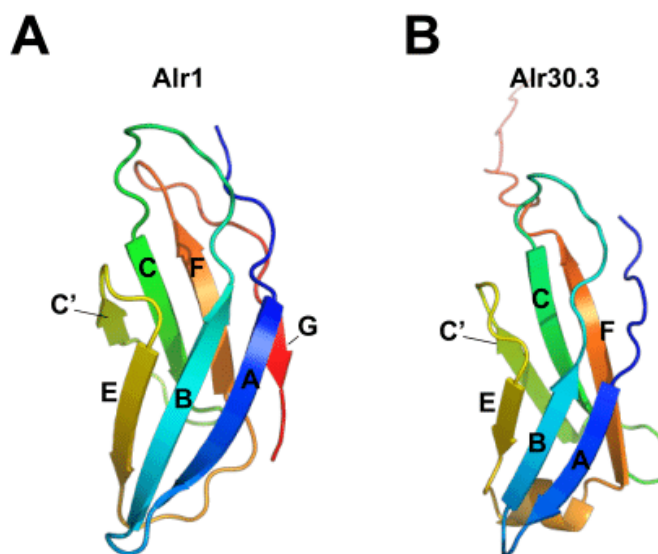
To further explore the similarity of the ECS to FN3 domains, I investigated whether they shared the same number and arrangement of  $\beta$ -strands. FN3 domains have an Ig-like fold composed of seven  $\beta$ -strands. The strands are labeled A, B, C, C', E, F, and G, analogous to the naming scheme for the C2-set immunoglobulin domains (Leahy et al., 1992; Halaby et al., 1999). In both FN3 and C2-set domains, strands A, B, and E form one  $\beta$ -sheet, while strands G, F, C, and C' form the other (Figure 35). This arrangement differs from C1-set domains, in which strands A, B, E, and D form one sheet and strands G, F, and C form the other (Halaby et al., 1999; Bodelon et al., 2013). Because the fourth strand “switches” sheets, it is labeled D in C1-set domains, and C' in C2 and FN3 domains (Halaby et al., 1999).



**Figure 35. Topology of  $\beta$ -strands in Fn3 domains, C1-set Ig domains, and C2-set Ig domains.**

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'.

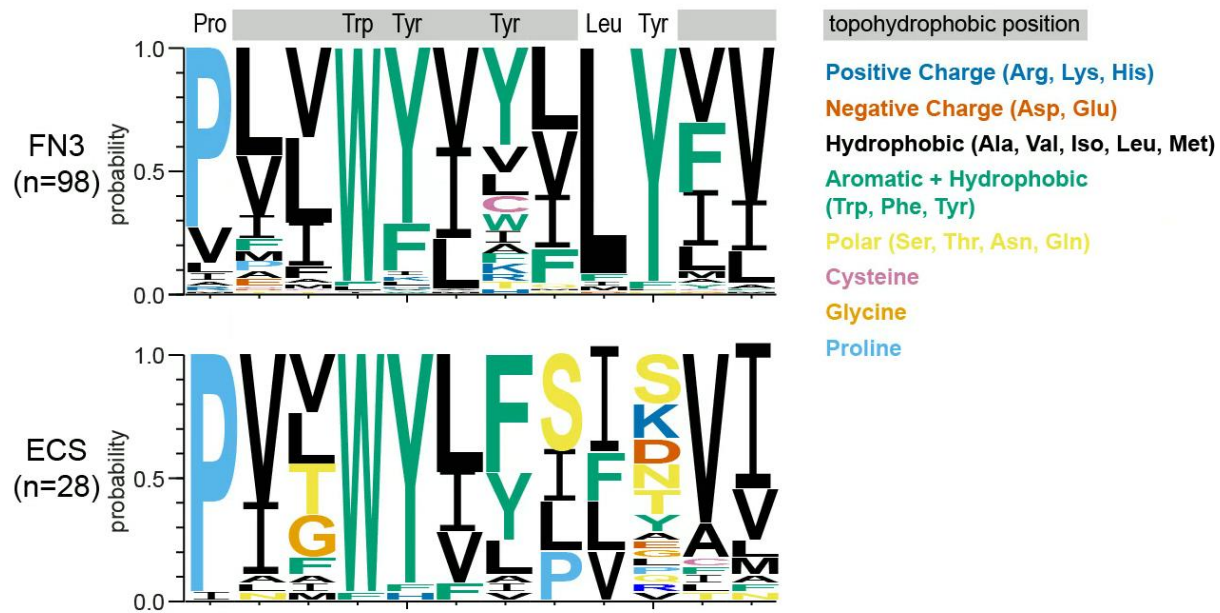
Next, I inspected each ECS model in PyMOL and found that the carbon backbone traced the path of a C2/FN3 Greek key (e.g. Figure 36). I then assigned letters to  $\beta$ -strands predicted by STRIDE (Figure S12) and found that 28/29 models had seven  $\beta$ -strands. In one model, the G strand was not predicted by STRIDE (Figure 36B). When inspected, the residues which would have been expected to contain the G strand appeared as a flexible tail rather than forming a strand alongside the domain. Thus, nearly all ECS structural models had a FN3-like Greek key topology.



**Figure 36. Predicted topology of  $\beta$ -strands in the Alr1 and Alr30.3 ECS folds.**

Alr1 conforms to an Fn3-like arrangement. Alr30.3 lacks the predicted G strand possibly due to the flexibility of that region in the model.

The primary amino acid sequences of FN3 domains have six conserved amino acids (Leahy et al., 1992; Halaby et al., 1999). To determine whether ECS sequences had these residues, I aligned them to FN3 sequences from the seed alignment of the Pfam FN3 profile (pf00041.23). Across the entire alignment, the ECS sequences aligned best to FN3 domains toward their N-terminus. 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 (Figure 37). 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.



**Figure 37. Sequence logo of amino acid frequencies in canonical Fn3 and the Alr ECS.**

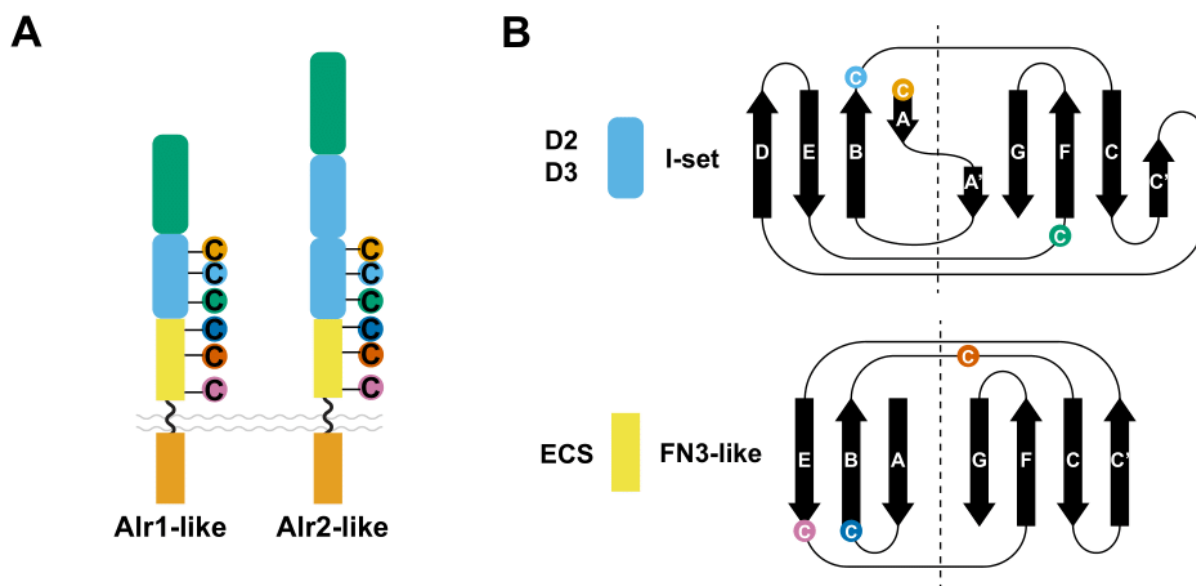
Sequence logo showing residues at conserved positions within Fn3 domains (top) and the ECS fold (bottom).

Fn3 domains also have eight ‘topohydrophobic’ positions (i.e., positions usually occupied by VILFMWY residues) (Halaby et al., 1999). These are mainly located within the  $\beta$ -strands that make up the core of the fold. Two of them are the conserved tyrosines in strand C, as described above. Each of the remaining six positions was occupied by a VILFMWY residue more than 50% of the time, although ECS sequences violated this rule more often than Fn3 sequences (Figure 37).

Taken together, these data show that the primary amino acid sequence is more similar to Fn3 domains than Ig domains and suggest the ECS adopts an immunoglobulin fold with an Fn3-like topology.

### 3.4.4 The membrane proximal domains of Alr proteins have six conserved cysteines

While investigating protein alignments of the Alr domains, I identified six cysteine residues that were conserved across all sequences. Three were in the ECS and three in the domain that immediately preceded it (Figure 38A). Within the ECS, the first and third cysteine were at the start of strand B and the end of strand E (Figure 38B). In the structural models, these residues were near each other and might form a disulfide bond in the native protein. The remaining cysteine was in the BC loop. Within the D2 (or D3) domains, the first two cysteines were found at the beginning of  $\beta$ -strand A and at the end of  $\beta$ -strand B, and were also near each other, raising the possibility that they might form a disulfide bond. The remaining cysteine was found just before strand F.

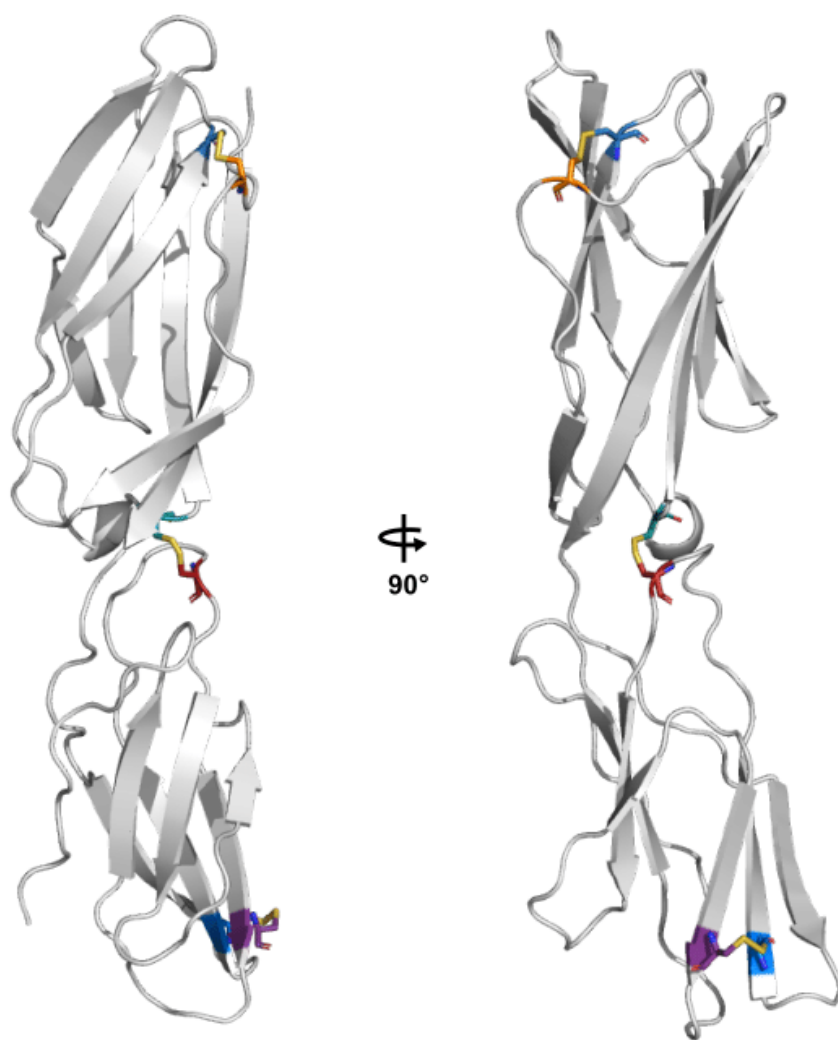


**Figure 38. Invariant cysteines in domains 2, 3, and the ECS fold.**

A) Occurrence of invariant cysteine residues found in Alr proteins with an Alr1-like or Alr2-like domain architecture. The cysteines are always found in the two membrane-proximal folds. B) Position of invariant cysteines in domains 2, 3, and the ECS fold.

To determine whether the two ‘unpaired’ cysteines might interact with each other, I used AlphaFold to predict structures for the tandem D2-ECS-trimmed or D3-ECS-trimmed domains. Both cysteine pairs and the two unpaired cysteines were predicted to form intramolecular disulfide bonds in 25/28 of the predictions when visually inspected (Figure 39). In the remaining three predictions, the cysteine pair in the ECS was measured to be slightly more distant (2.4 Å) than the average disulfide bond (2.05 Å) but within the 3.0 Å cutoff (M. Sun et al., 2017). Thus, it appears likely that all invariant cysteine residues may participate in intra-molecular disulfide bonds based on these predictions.



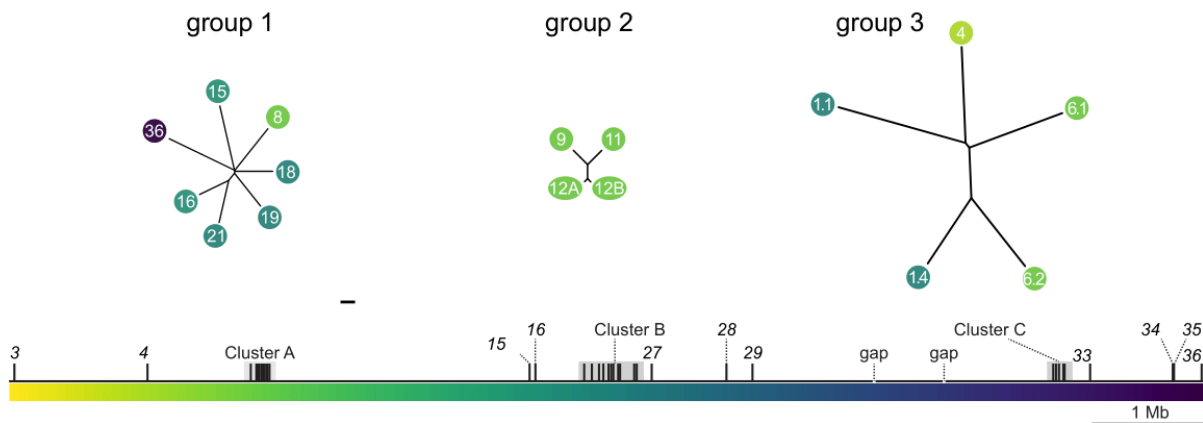


**Figure 39. Model of the invariant cysteines in Alr1 domain 2 through the ECS fold.**

Predicted location of invariant cysteines in the membrane-proximal folds shown in Alr1. Structural predictions of D2-ECS (or D3-ECS) were made with AlphaFold. The cysteines are highlighted in colors corresponding to Figure 38B.

### 3.4.5 The cytoplasmic tails of many Alr proteins contain ITAM or ITIM motifs

The Alr cytoplasmic tails were too diverse to be included in a single alignment. Therefore, I used CD-HIT to cluster them by sequence identity. This placed half of the cytoplasmic tails (16/32) into three groups (Figure 40). Group members could then be aligned to one another. Of the remaining 16, three were short (<14 aa) and the rest could not be grouped with another sequence. Based on this, the cytoplasmic tails of Alr proteins are apparently more divergent from one another than their extracellular domains.

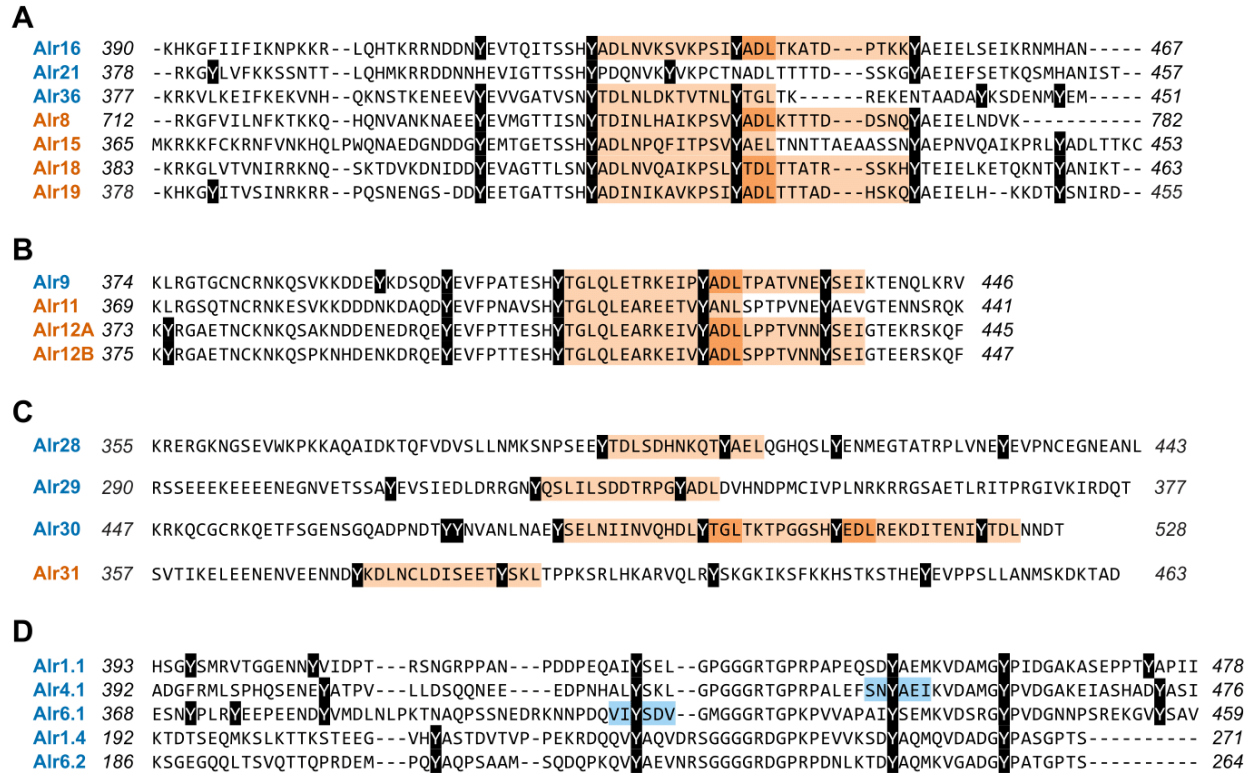


**Figure 40. Alr cytoplasmic tails are diverse.**

Cytoplasmic tails grouped by CD-HIT at N% 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.

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 (Figure 41A-C). ITAMs, which have a consensus

sequence of Yxx[I/L]x<sub>(6-9)</sub>Yxx[L/I] (Murphy et al., 2008), are found in receptors that activate immune responses in vertebrates (Bezbradica & Medzhitov, 2012) and stimulate phagocytosis of damaged cells in *Drosophila* (Ziegenfuss et al., 2008). 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, which have a consensus sequence of [I/L/V/S]xYxx[I/V/L] (Barrow & Trowsdale, 2006), in two Alr tails, both in group 1 (Figure 41D). Thus, many Alr gene family members bear motifs found in receptors that regulate the recognition of non-self in other animals.

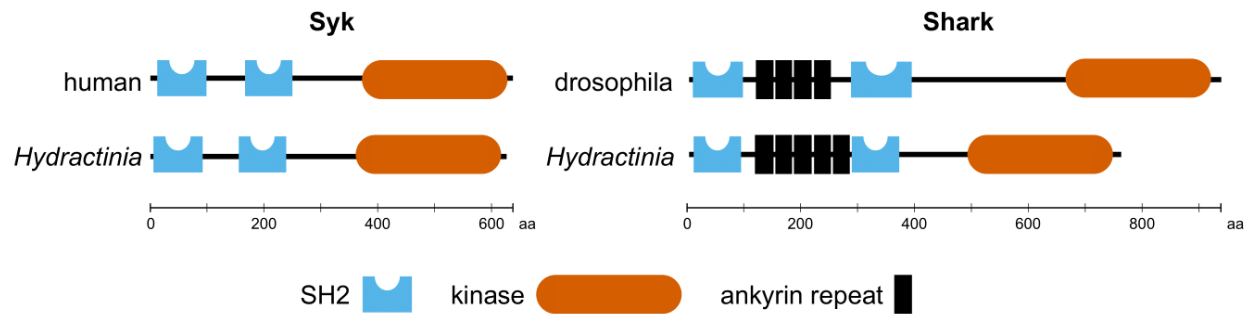


**Figure 41. Sequence analysis of cytoplasmic tails and their ITAM and ITIM motifs.**

A) Alignment of group 1 cytoplasmic tails. B) Alignment of group 2 cytoplasmic tails. C) ITAMs in ungrouped cytoplasmic tails. In (A-C), bona fide gene names are in blue, putative gene names are in orange. ITAMs have an orange background. Darker orange indicates overlapping ITAMs. D) Alignment of group 3 cytoplasmic tails. ITIMs have a blue background. All tyrosines have a black background. In (A-D) the alignment is truncated to highlight the ITIM/ITAM motifs, and all tyrosines are shown with a black background.

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 a set of ankyrin repeats is found between the 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 (Figure 42). 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). Thus, the *Hydractinia* genome encodes ITAM-bearing receptor-like proteins and orthologs of two kinases that potentially bind them.



**Figure 42.** Comparison of human Syk and *Drosophila* Shark to *Hydractinia* Syk and Shark.

### 3.5 Discussion

Alr domains 1-3 were most similar to members of Ig domains, despite the lack of several motifs usually found in Ig domains. The most important of these, from the standpoint of differentiating Ig domains from other immunoglobulin-like folds, is the C-W-C (or “pin” motif). While I could not find any reports of Ig domains lacking this motif entirely, the analyses of the primary, secondary, and predicted tertiary structures of domains 1-3 consistently showed them to be most similar to Ig domains and not other proteins with immunoglobulin-like folds. I think it is unlikely that this level of overall similarity to Ig domains evolved via convergent evolution. Therefore, I propose domains 1-3 are novel members of the IgSF.

With respect to the evolutionary history of the V-set and I-set Ig domains, the data are consistent with two scenarios. In the first, the cnidarian-bilaterian common ancestor had distinct V-set-like and I-set-like domains. These domains could have then followed different evolutionary

trajectories to arrive at the current sequences of extant Alr, V-set, and I-set domains. The alternative scenario is that the cnidarian-bilaterian common ancestor had an Ig domain that was not distinctly V-set-like or I-set-like. This Ig domain then evolved into the current Alr domains, V-set domains, and I-set domains. The similarities between these domains would be the result of convergent evolution. Testing these hypotheses will require analyzing Ig domains from additional metazoan genomes and obtaining experimental structures for at least some of them.

My analyses have also revealed that part of the ECS is likely to adopt an immunoglobulin-like  $\beta$  sandwich. The predicted secondary and tertiary structure of this fold was most similar to Fn3 domains. However, its primary amino acid sequence differed substantially from Fn3 domains, especially over its C-terminal region. Indeed, search algorithms that rely on primary amino acid sequences were unable to detect any similarity between Fn3 and ECS sequences. Therefore, it is unclear whether the ECS fold is homologous to members of the Fn3 superfamily or represents a new type of immunoglobulin-like fold.

The domain architecture of Alr proteins suggests that they could play roles in extracellular protein-protein interactions, signaling, or adhesion. Tandem Ig domains are commonly found in cell adhesion molecules, proteins involved in cell-to-cell communication, and immune receptors. An adhesive function would be consistent with that already described for Alr1 and Alr2 (Karadge et al., 2015; Huene et al., 2021). This property might be expected for any Alr that functions as an additional allodeterminant. In this context, the membrane-proximal ECS-fold might correctly position the Alr Ig domains to bind their ligands. Moreover, if this domain had Fn3-like properties, it might also contribute some elasticity to the protein. It has been suggested that Fn3 domains unfold and refold *in vivo*, acting like molecular springs (Smith et al., 2007; Kubow et al., 2015) but this hypothesis is controversial (Erickson, 2017).

Another clue to Alr function comes from the six invariant cysteine residues in their membrane-proximal extracellular domains. This conservation is especially striking because there is only one other invariant position in the Alr Ig domains — a proline in domain 2. Within each domain (I-set-like or ECS fold), each of the cysteine pairs appear to be close enough to form disulfide bonds based on the structural predictions. This disulfide bonds would provide stability to their respective domains. If there exist alternative conformations for these two domains which prevent disulfide bond formation, it is possible that they could participate in the formation of homo- or hetero-dimers. In particular, if the orientation of domain 2 or domain 3 to the ECS were rotated, the singular cysteines in domain 2 or domain 3 and the ECS may then be surface-exposed, thereby allowing them to participate in the formation of homo- or hetero-dimers. Such dimerization, if it affects binding specificities, could add another layer of complexity to how Alr1 and Alr2 discriminate self from non-self.

The ITAM motifs in some Alr cytoplasmic tails are also potentially significant. The dual tyrosines in ITAM motifs are typically phosphorylated by Src family kinases (SFKs) and then bound by Syk-like kinases (Mócsai et al., 2010). SFK-ITAM-Syk signaling often occurs within cells that respond to pathogenic non-self or self tissues that are damaged or unwanted. For instance, an SFK-ITAM-Syk signaling module enables glial cells to phagocytose apoptotic neurons in *Drosophila* (Ziegenfuss et al., 2008) and appears to promote responses to bacteria in phagocytic cells in oysters (J. Sun et al., 2020). It can also induce inflammation in epithelial cells in vertebrates (Hoft et al., 2020), where it also plays a well-characterized and essential role in the activation of immune cells (Futosi & Mócsai, 2016; Au-yeung et al., 2017). Alr proteins with ITAMs might have similar functions.

Could SFK-ITAM-Syk mediated signaling play a role in allorecognition responses? One possibility is that Alr proteins with ITAM motifs activate rejection responses when they bind relatively invariant *Hydractinia*-specific ligands on opposing tissues. This rejection response could then be inhibited if polymorphic allodeterminants (Alr1, Alr2, and likely others) bind a compatible ligand. At present, this model is only supported by two seemingly disparate observations. First, *Hydractinia* colonies only mount allorecognition responses in response to specific hydroids (and always to other *Hydractinia*), suggesting they can identify the type of tissue that they encounter, possibly via invariant cell surface receptors. Second, the initial stages of rejection and fusion are morphologically indistinguishable. In both responses, nematocytes migrate to the point of contact and arrange their nematocysts as batteries that are oriented toward their opponent. In a rejection, these batteries then fire, but in a fusion, the nematocytes migrate away from the zone of contact as the tissues merge. This suggests that rejection is the default allorecognition response in *Hydractinia*. Alrs with ITAMs might activate rejection, which is then inhibited by homophilic binding between proteins encoded by compatible *Alr1* or *Alr2* alleles. In the mammalian immune system, this inhibitory role is filled by receptors with cytoplasmic ITIMs. Neither Alr1 nor Alr2 have an ITIM, but Alr4 and Alr6 do. 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.



## **3.6 Methods**

### **3.6.1 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). For HMMER sequence homology, hmmscan was used to query each domain sequence against the Pfam database. 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  $1 \times 10^{-3}$  for inclusion. HHpred was then used to search the SCOPe70\_2.07 database.

### **3.6.2 Alr sequence comparisons**

Alignments between Alr proteins were performed using MAFFT (Katoh & Toh, 2010) as implemented in Jalview (Waterhouse et al., 2009). 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. Neighbor joining trees were constructed in Jalview using the BLOSUM62 scoring matrix. Trees were visualized in iTOL (Letunic & Bork, 2019), exported as scaled vector graphics files, and annotated in Adobe Illustrator.

### 3.6.3 Structural predictions and visualization

Structural predictions were performed using AlphaFold (Jumper et al., 2021), as implemented in Colabfold (Mirdita et al., 2021). Each *Alr* sequence and its corresponding HHblits MSA query results were submitted as an a3m files. The top structural model of each prediction was submitted to the DALI server (<http://ekhidna2.biocenter.helsinki.fi/dali/>, Holm, 2020) to identify structures from the PDB (rcsb.org; Berman et al., 2000) with the same fold. The top hit was downloaded and if necessary modified to contain only the relevant chain for structural alignment. Each AlphaFold prediction was then submitted with its identified PDB structure to TMalign (<https://zhanggroup.org/TM-align/>, Zhang & Skolnick, 2005) to obtain the TM-score of the alignment. Secondary structure was predicted with STRIDE (Heinig & Frishman, 2004). Structural models were visualized in PyMOL 2.0 (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.).

## 4.0 New self-identities evolve via point mutation in an invertebrate allorecognition gene

### 4.1 Foreword

This chapter is adapted from a work previously published in iScience in which I am first author:

Huene, A.L., Chen, T., Nicotra, M.L., 2021. New binding specificities evolve via point mutation in an invertebrate allorecognition gene. iScience 24, 1–24. <https://doi.org/10.1016/j.isci.2021.102811>

### 4.2 Summary

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 non-self by binding in trans to other *Alr2* proteins with identical or similar sequences. Here, I focused on the N-terminal domain of *Alr2*, which can determine its binding specificity. I 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. I also show that one side of the domain has experienced positive selection and likely forms the binding interface. These 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.

### 4.3 Introduction

*Alr1* and *Alr2* both exhibit highly specific interactions between identical or near identical isoforms (Karadge et al., 2015). These results combined with the extreme levels of polymorphism observed in *Alr1* and *Alr2* in nature (Rosa et al., 2010; Gloria-Soria et al., 2012) 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 & 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.

Under negative frequency-dependent selection, alleles become fitter as they become rarer. 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 those of other

alleles (A. D. Richman & Kohn, 2000). These dynamics can maintain tens to hundreds of self-recognition alleles in a population (Casselton & 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.

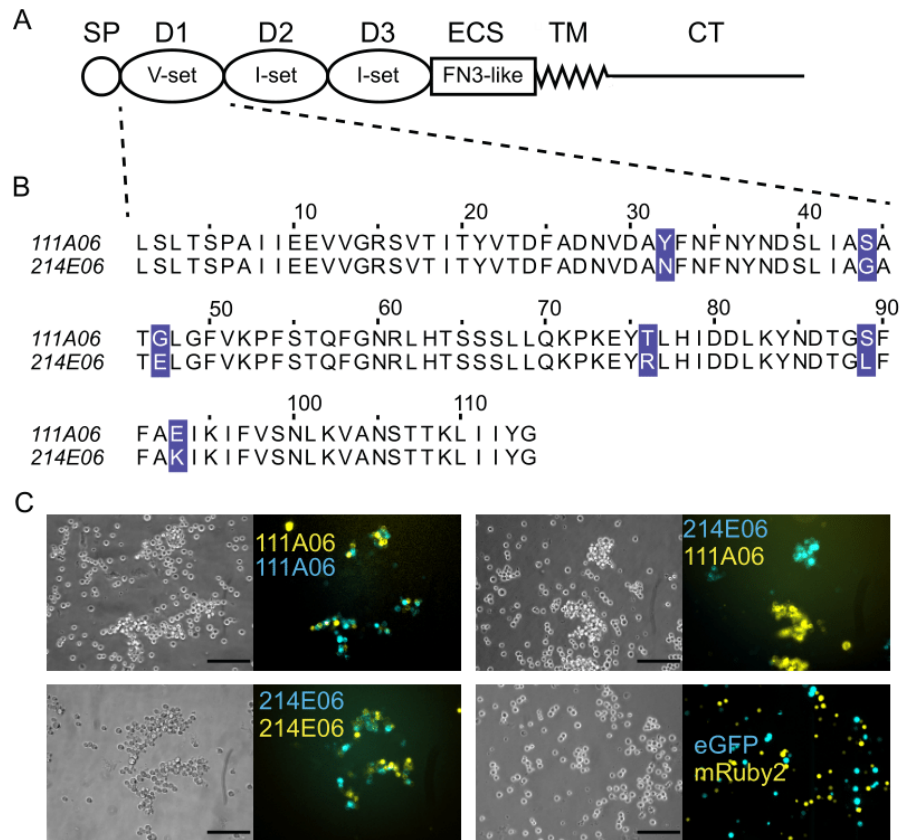
How novel *Alr1* and *Alr2* alleles are generated has been assumed to be through random mutations that are then subjected to negative frequency-dependent selection. This raises the question of whether point mutations, by themselves, can generate alleles with novel homophilic binding specificities and,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

Here, I 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, I identified a clade of five domain 1 sequences encoding isoforms that differed by six or fewer amino acids. Then, I used ancestral sequence reconstruction and in vitro binding assays to determine the evolutionary history of the clade. My 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, I 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.

## 4.4 Results

### 4.4.1 Point mutations in domain 1 can create new binding specificities

I searched a dataset of full-length, naturally occurring *Alr2* alleles (Nicotra et al., 2009; Gloria-Soria et al., 2012), and identified two (111A06 and 214E06) 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 (Figure 43A, B). Using cell aggregation assays, I found that each isoform bound to itself across opposing cell membranes but did not bind to the other (Figure 43C). These results were performed in triplicate and produced repeatable results (Supplemental Figure 1). I therefore sought to identify the amino acid differences that prevented them from binding to each other.

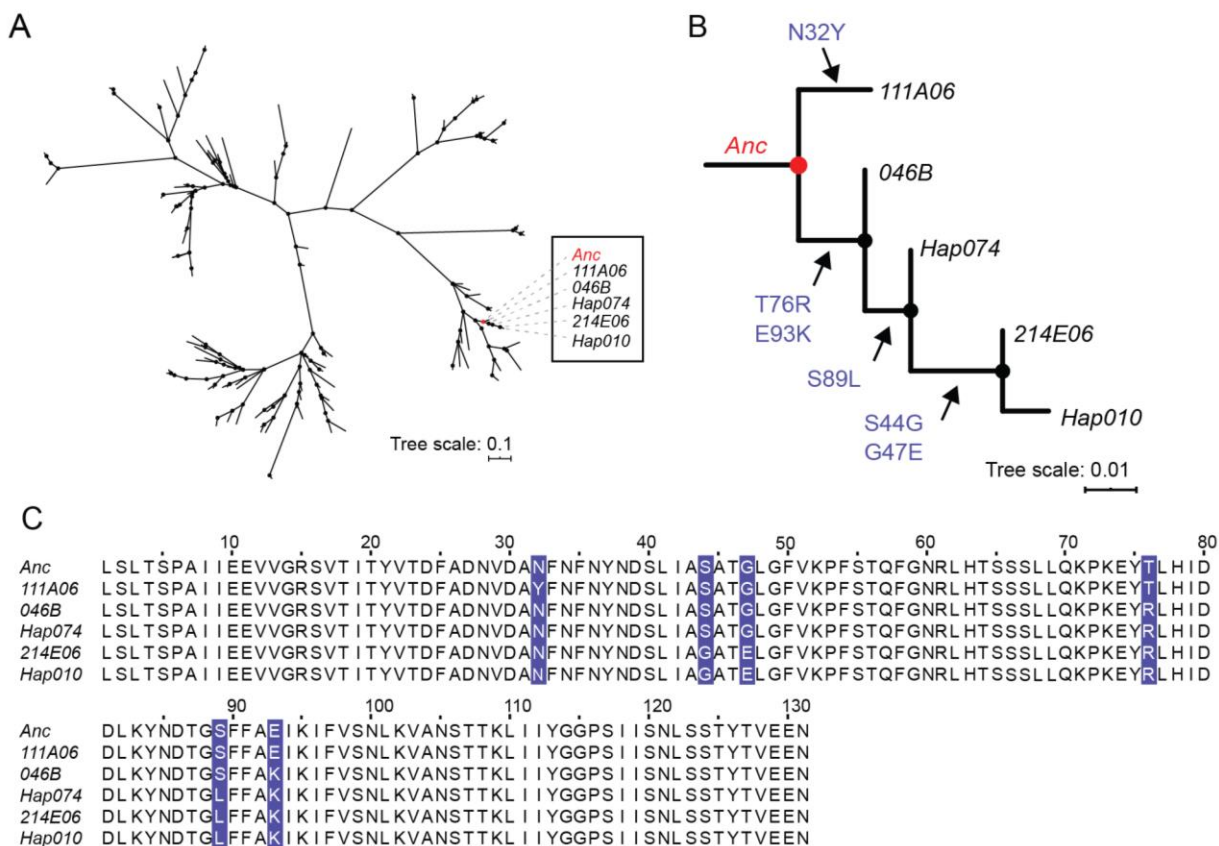


**Figure 43 Isoform-specific, homophilic binding of Alr2 isoforms.**

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).

Each amino acid difference between 111A06 and 214E06 is the result of one point mutation. To reconstruct the evolutionary history of these mutations, I created a phylogeny of all known domain 1 coding sequences (Figure 44A). 111A06 and 214E06 were located in a clade with three additional sequences (Figure 44B). I 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 44C). Because 214E06 and Hap010 differed only by a single synonymous mutation, I used 214E06 to represent their shared amino acid sequence.



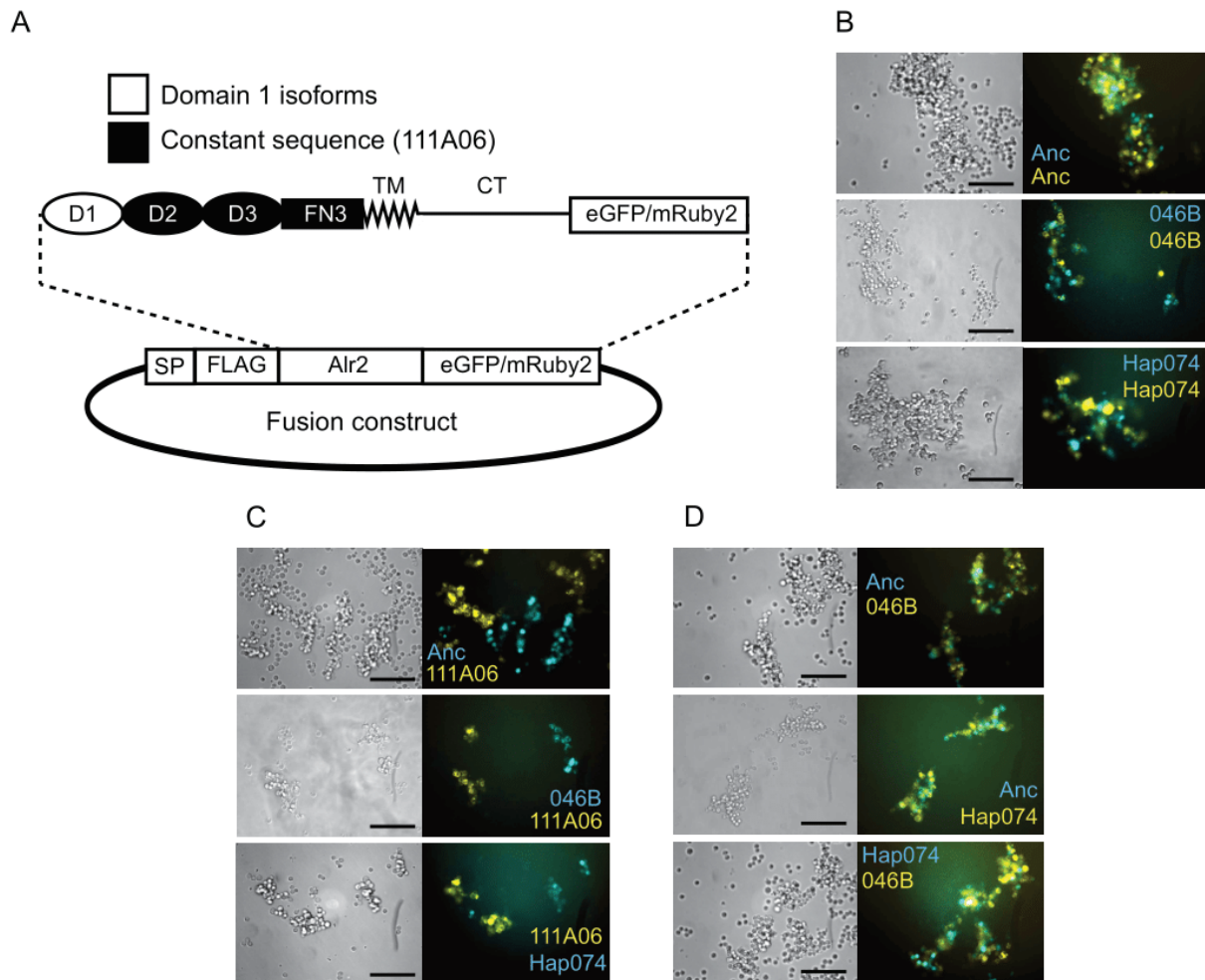
**Figure 44 Relationship of five naturally occurring *Atr2* alleles.**

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

To determine the binding specificity of the domain 1 isoforms encoded by these sequences, I expressed each as a fusion to domain 2 through the cytoplasmic tail of the 111A06 isoform, with a C-terminal fluorescent protein tag (Figure 45A). The resulting isoforms were tested against themselves and each other in cell aggregation assays. Each isoform, including the predicted ancestor, Anc, caused cells to form multicellular aggregates, indicating it was capable of



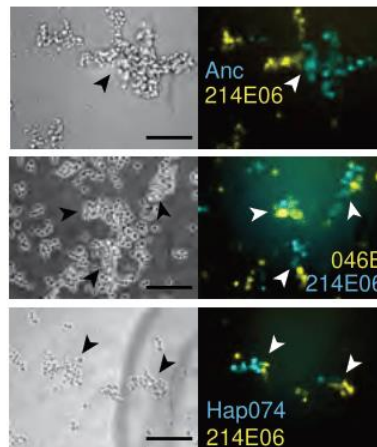
homophilic binding (Figure 45B). In pairwise assays, 111A06 did not form mixed aggregates with any isoform, indicating it had a unique binding specificity within the clade (Figure 45C). In contrast, Anc, 046B, and Hap074 all formed mixed aggregates with each other, indicating a shared binding specificity (Figure 45D). These results were performed in triplicate and produced repeatable results (Supplemental Figure 1).



**Figure 45. Plasmid template used in cell aggregation assay and assay results.**

A) Plasmid map Alr2 fusion proteins (TM = transmembrane domain, CT = cytoplasmic tail). B-D) Representative images of cell aggregation assays (See Supplemental Figure 1 for replicates). B) Anc, 046B, and Hap074 against themselves. C) Anc, 046B, and Hap074 against 111A06. D) All pairwise combinations of Anc, 046B, and Hap074.

In assays that paired 214E06 with Anc, 046B, or Hap074, I observed single-color aggregates, some of which appeared to adhere to aggregates of a different color (Figure 46, arrowheads). These semi-mixed aggregates were repeatable (Supplemental Figure 1A) 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. Semi-mixed 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, I 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. These results were performed in triplicate and produced repeatable results (Supplemental Figure 1A).



**Figure 46. Anc, 046B, and Hap074 versus 214E06.**

Arrowheads point to semi-mixed aggregates (See Supplemental Figure 1A for replicates).

These results are consistent with the following evolutionary history (Figure 47). 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 47, dotted lines). The result is a clade in which I can discern three binding specificities, one of which arose via a single point mutation.

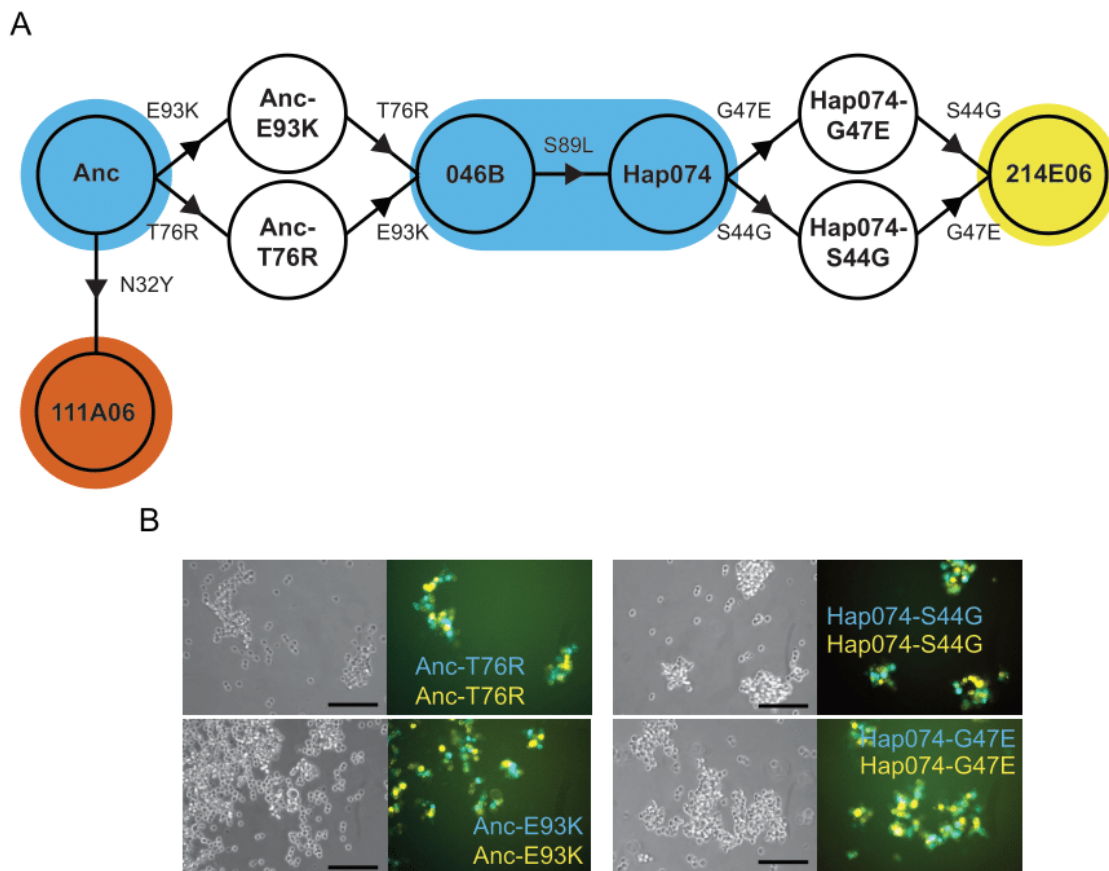


**Figure 47. Node network of isoforms colored by binding specificity.**

Triangles indicate the hypothesized direction of mutation from Anc. Green dotted lines indicate weaker heterophilic interactions.

#### **4.4.2 New homophilic specificities can evolve via less restricted intermediates**

Within the phylogeny, two pairs of mutations occurred within single branches (Figure 44B), 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, I re-created each one (Figure 48A) and tested it in cell aggregation assays. I found each intermediate could bind homophilically (Figure 48B), thus ruling out the possibility that there were non-functional intermediates in the clade.

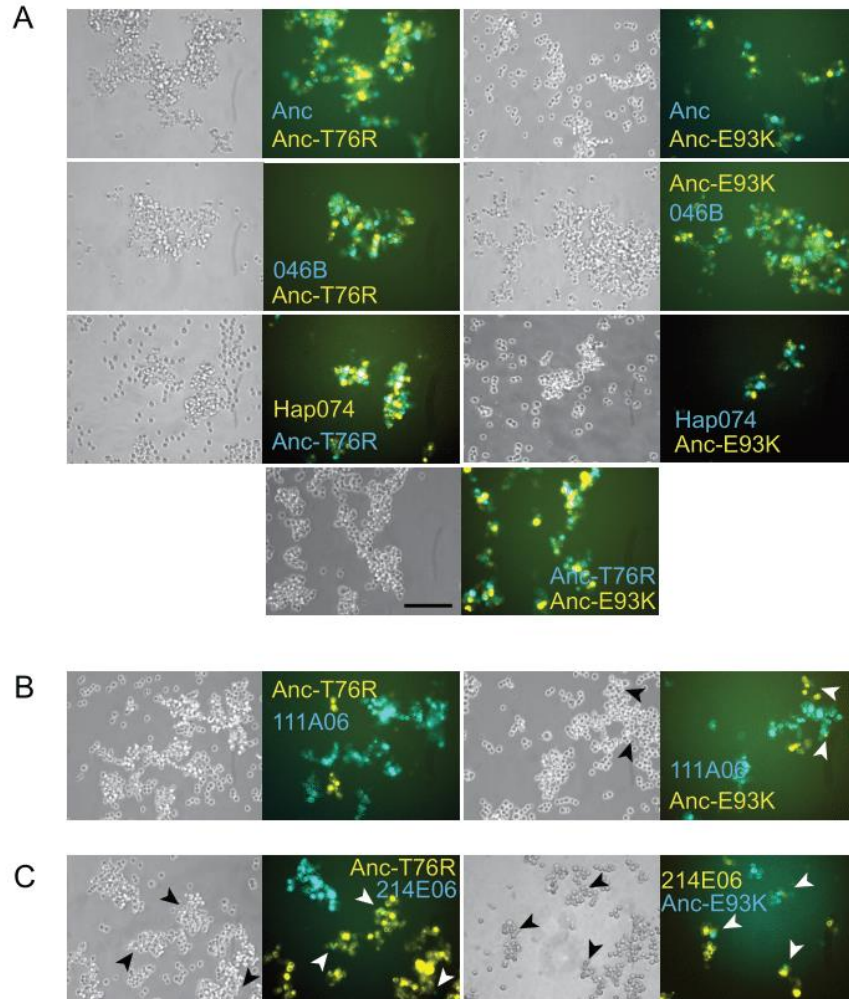


**Figure 48. Single-step domain 1 mutants are capable of homophilic binding.**

A) Expanded node network including hypothesized single-step mutants between Anc and 046B, Hap074 and 214E06. B) Representative images of single-step mutants tested against themselves. (See Supplemental Figure 1 for replicates)

I 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 49A). Assays pairing Anc-T76R with 111A06 resulted in single color aggregates (Figure 49B), but those pairing Anc-E93K with 111A06 resulted in a few semi-mixed aggregates (Figure 49B, arrowheads, Supplemental Figure 1). Both mutants also formed semi-mixed aggregates when paired with

214E06 (Figure 49C, Supplemental Figure 1). Thus, evolution from Anc to 046B is unlikely to have involved a significant change in binding specificity.



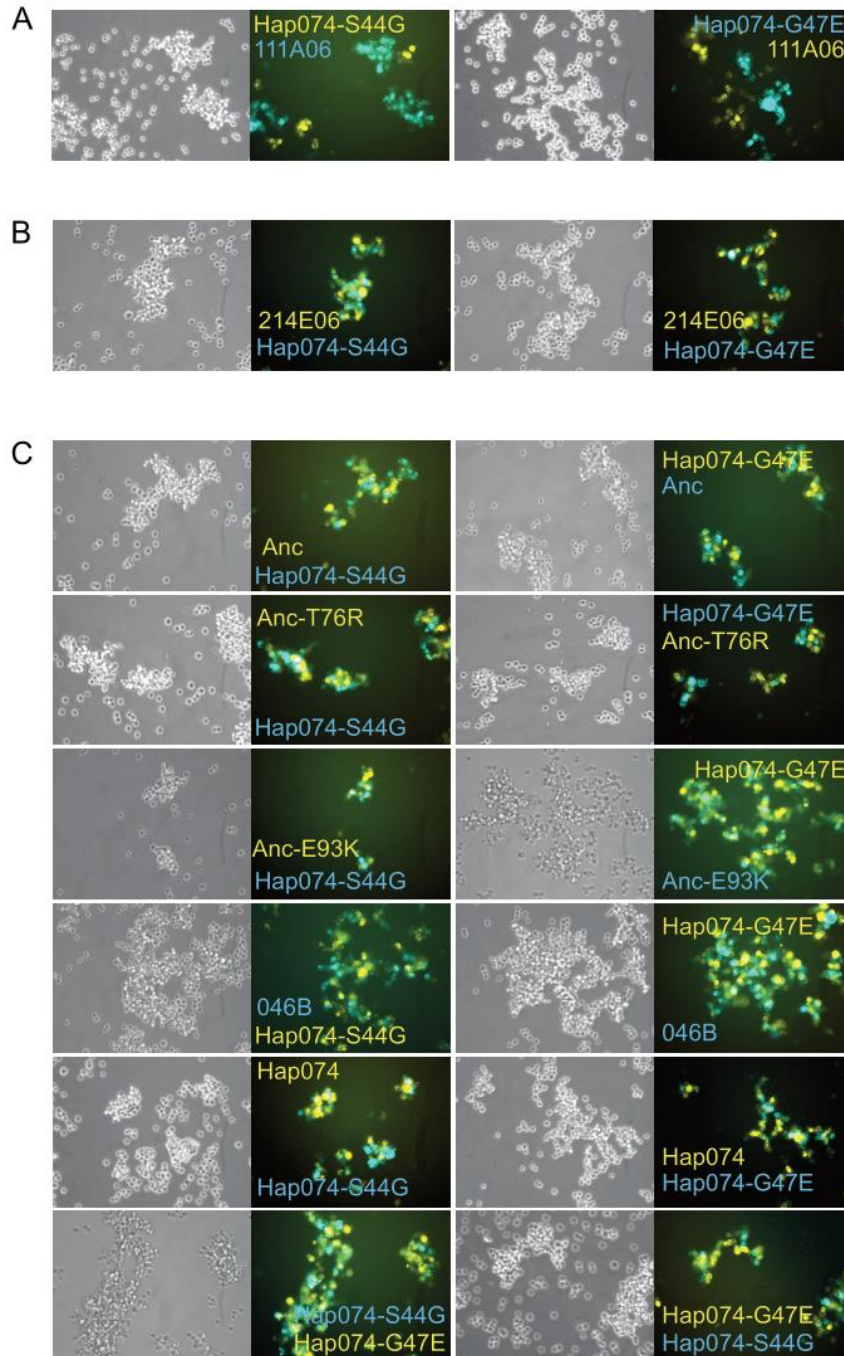
**Figure 49. Anc-T76R and Anc-E93K cell aggregation assay pairwise results.**

A-C) Representative images of cell aggregation assays. A) Anc-T76R and Anc-E93K tested against Anc, 046B, Hap074. B) Anc-T76R versus 111A06 (left) and Anc-E93K versus 111A06 (right). Semi-mixed aggregates indicated with arrowheads (See also Fig S1A). C) Anc-T76R and Anc-E93K versus 214E06 (See Supplemental Figure 1 for replicates).

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

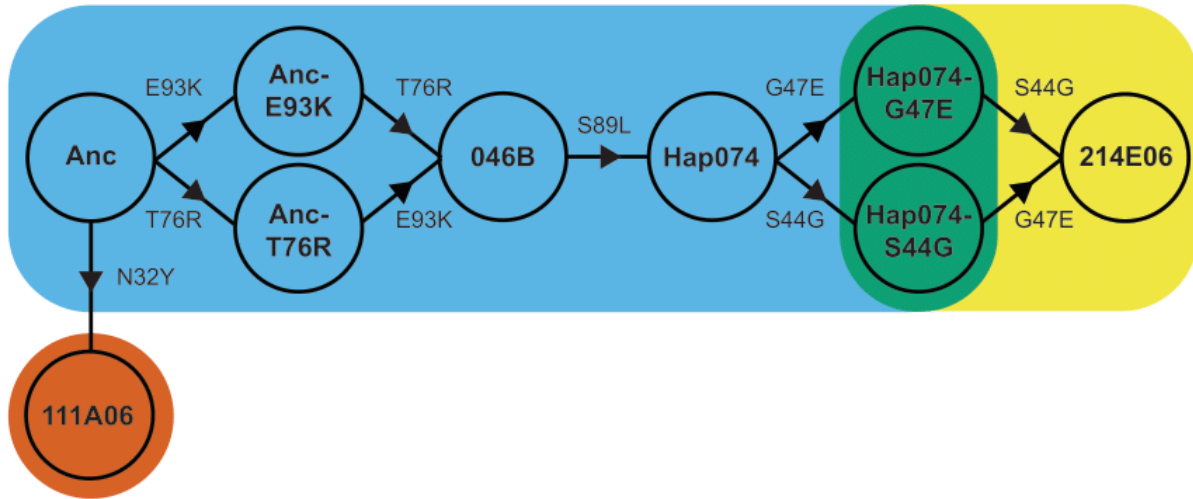
aggregates with 111A06 (Figure 50A) but did form mixed aggregates with 214E06 (Figure 50B) and all other ancestral sequences (Figure 50C). I did not observe semi-mixed aggregates in any assay. These results suggest the first mutation on the path from Hap074 to 214E06, either S44G or G47E, created a sequence that could still bind Hap074. 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 (Figure 51).





**Figure 50. Hap074-S44G and Hap074-G47E cell aggregation assay results.**

A-C) Representative images of cell aggregation assays (See Supplemental Figure 1 for replicates). A) Hap074-S44G and Hap074-G47E versus 111A06. B) Hap074-S44G and Hap074-G47E versus 214E06. C) Hap074-S44G and Hap074-G47E versus all remaining isoforms.



**Figure 51. Domain 1 isoforms can evolve via intermediates with broadened specificity.**

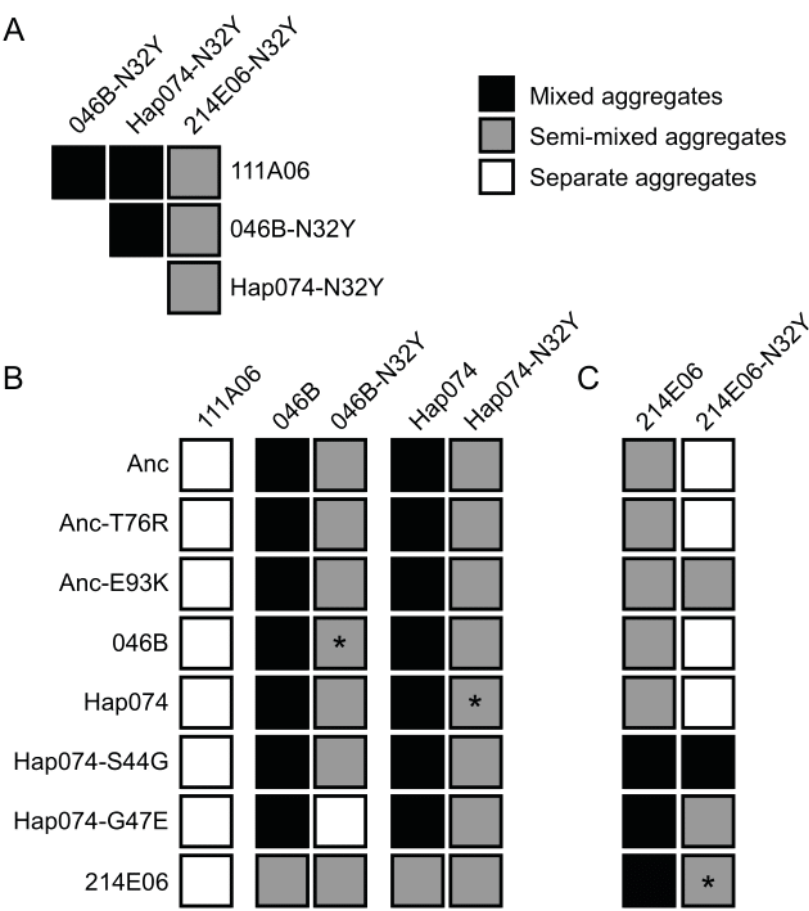
Expanded node network including hypothesized single-step mutants between Anc and 046B, Hap074 and 214E06 with their binding specificity highlighted.

#### 4.4.3 The N32Y mutation preserves homophilic binding and alters specificity

Isoform 111A06 evolved when position 32 mutated from Asn to Tyr in Anc. I 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, I generated 046B-N32Y and Hap074-N32Y. In assays with themselves, each formed mixed aggregates, indicating the mutation did not disrupt homophilic binding (Supplemental Figure 2D). 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 (Figure 52A and Supplemental Figure 2G). In pairwise assays with their immediate ancestors, however, the mutants formed semi-mixed aggregates (Figure 52B, asterisks, and Supplemental Figure 2). This indicated each could still bind its ancestor, albeit more weakly than it did itself. Finally, I performed pairwise assays with the remaining isoforms in the clade.



This showed the mutants had different specificities than 111A06, 046B, or Hap074 (Figure 52B and Supplemental Figure 2). In sum, the N32Y altered the specificities of 046B and Hap074 but did not generate daughter sequences with the same specificity as 111A06.



**Figure 52. Effects of N32Y mutation on binding specificity.**

A) Results of assays between N32Y mutants and 111A06 (See also Supplemental Figure 2G,H). 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.

I next tested whether the N32Y mutation would alter the specificity of 214E06, the remaining domain 1 isoform known to exist in nature. I generated 214E06-N32Y and found it formed mixed aggregates with itself (Supplemental Figure 3D), indicating it was able to bind

homophilically to itself. It also formed semi-mixed aggregates with 214E06 (Figure 52C, asterisk, and Supplemental Figure 3F), indicating a reduced binding affinity for its immediate ancestor compared to itself. However, 214E06-N32Y only formed semi-mixed aggregates with 111A06, 046B-N32Y, and Hap074-N32Y (Figure 52A). 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 (Figure 52C and Supplemental Figure 3). The effect of the N32Y mutation thus depends on the sequence context in which it occurs.

#### **4.4.4 Structural and evolutionary analyses suggest a potential binding interface**

In this study, the change of three residues correlated with the change in binding specificity of domain 1 (N32Y, S44G, and G47E). To investigate how these mutations might affect the tertiary structure of domain 1—and thus its binding specificity—I used AlphaFold to predict their structures. All were predicted to fold like V-set Ig-domains, which was consistent with previous work (Nicotra et al., 2009) and my structural analysis in 3.4.1. All of the structural predictions had average (model-wide) pLDDT scores ranging from 93.7 to 95.1 (Table 10).

**Table 10. Predicted structural homology of Alr2 domain 1 isoforms.**

Isoform	AlphaFold <i>pLDDT</i> score <sup>a</sup>	DALI Top Structural Alignment					TM-align
		<i>PDB</i> accession	<i>Model</i>	<i>Z-score</i> <sup>b</sup>	<i>LALI</i> <sup>c</sup>	<i>RMSD</i> <sup>d</sup>	<i>TM-score</i> <sup>e</sup>
<i>Alr2 D1</i>							
111A06	95.1	<a href="#">7kqy-E</a>	Heavy-chain only human antibodies	15.2	108	1.9	0.81
Anc	94.6	<a href="#">7kqy-E</a>	Heavy-chain only human antibodies	15.2	108	1.9	0.81
Anc-T76R	94.7	<a href="#">6suz-H</a>	Major Prion Protein	15.3	107	1.8	0.81
Anc-E93K	94.6	<a href="#">6suz-H</a>	Major Prion Protein	15.2	108	1.9	0.81
046B	94.0	<a href="#">3oai-A</a>	Myelin protein P0	16.5	109	1.6	0.85
046B-N32Y	95.1	<a href="#">6suz-H</a>	Major Prion Protein	15.3	107	1.8	0.81
Hap074	93.9	<a href="#">3oai-A</a>	Myelin protein P0	16.5	109	1.6	0.85
Hap074-S44G	93.7	<a href="#">6suz-H</a>	Major Prion Protein	15.4	107	1.8	0.81
Hap074-G47E	94.5	<a href="#">3oai-A</a>	Myelin protein P0	16.5	109	1.6	0.85
Hap074-N32Y	94.0	<a href="#">3oai-A</a>	Myelin protein P0	16.4	109	1.6	0.85
214E06	94.7	<a href="#">7kqy-E</a>	Heavy-chain only human antibodies	15.3	108	1.9	0.81
214E06-N32Y	95.0	<a href="#">7kqy-E</a>	Heavy-chain only human antibodies	15.2	108	1.9	0.81

<sup>a</sup> predicted local-distance difference test score; values >80 are considered good models

<sup>d</sup> RMSD < 2 are considered reasonable models

<sup>b</sup> Z-score between 8-20 are considered probably homologous

<sup>e</sup> TM-scores > 0.5 indicate proteins with same general topology and are shaded green


<sup>c</sup> LALI = Number of equivalent residues considered in Z-score

Five of the six variant positions had pLDDT scores >92 in all models suggesting they have a high probability of having their side chains oriented correctly in the model (Table 11 and Figure 53). The remaining variant residue (G47/E47) had an pLDDT score that ranged between 75.0 and 92.2 (Table 11 and Figure 53). Its score was not correlated to its identity.

**Table 11. pLDDT scores of variant positions**

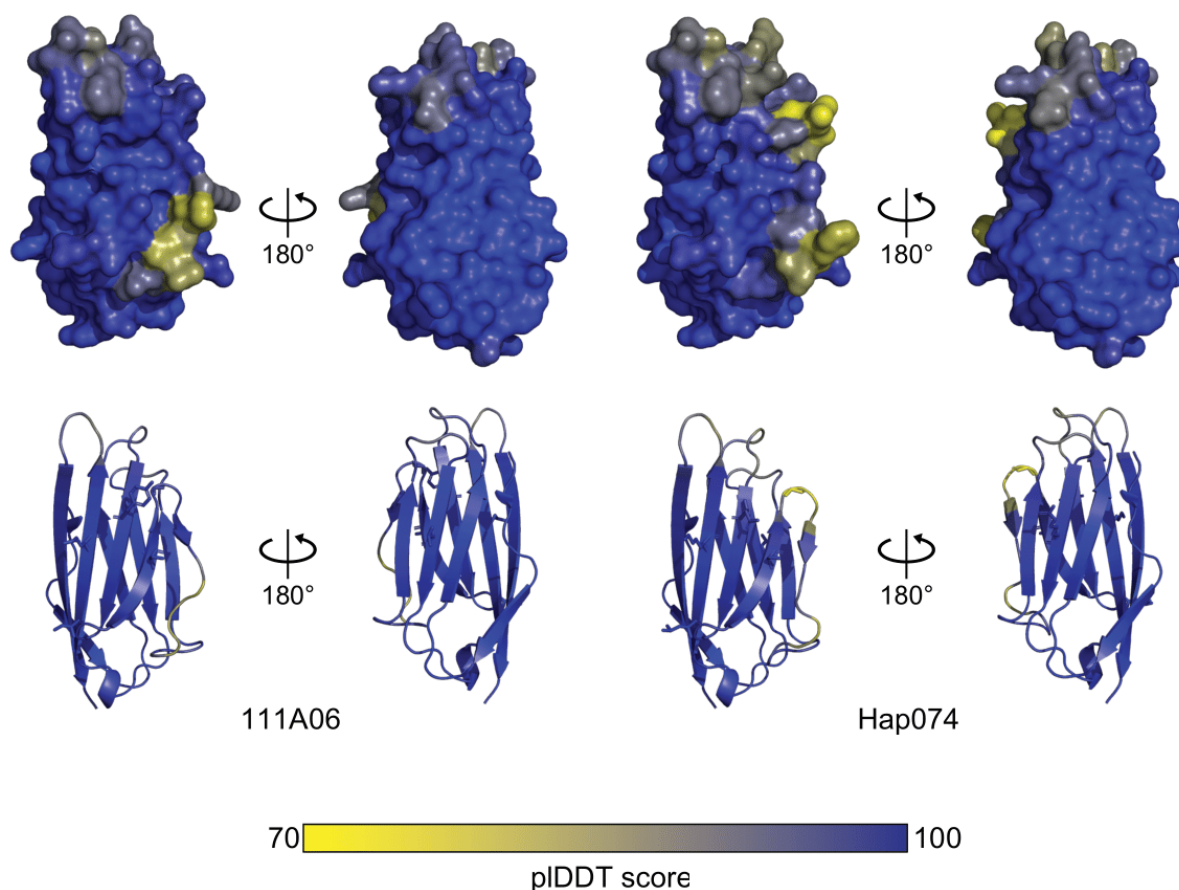
Color scale ranges from 70 to 100. Residues with pLDDT >90 are considered highly accurate and have their side chains oriented correctly 80% of the time. Residues with pLDDT >70 generally have their backbones predicted correctly.

Isoforms	pLDDT score						
	Model average	N32/ Y32	S44/ G44	G47/ E47	T76/ R76	S89/ L89	E93/ K93
111A06	95.1	96.2	97.0	90.9	98.0	98.4	98.5
Anc	94.6	94.9	96.3	88.6	97.7	98.4	98.3
Anc-T76R	94.7	95.3	96.7	89.9	98.0	98.4	98.4
Anc-E93K	94.6	95.0	96.3	88.7	97.6	98.3	98.3
046B	94.0	92.7	95.5	75.0	96.5	98.4	98.1
046B-N32Y	95.1	96.3	97.2	91.6	98.2	98.4	98.5
Hap074	93.9	92.6	95.4	74.6	96.3	98.5	98.0
Hap074-S44G	93.7	94.2	94.5	87.4	97.1	98.4	98.0
Hap074-G47E	94.5	93.4	96.5	80.6	97.1	98.5	98.2
Hap074-N32Y	94.0	93.2	95.7	75.1	96.2	98.5	98.0
214E06	94.7	95.4	95.8	92.2	97.8	98.2	98.0
214E06-N32Y	95.0	96.1	96.6	92.0	98.0	98.5	98.4
Average	94.5	94.6	96.1	85.5	97.4	98.4	98.2



70

100

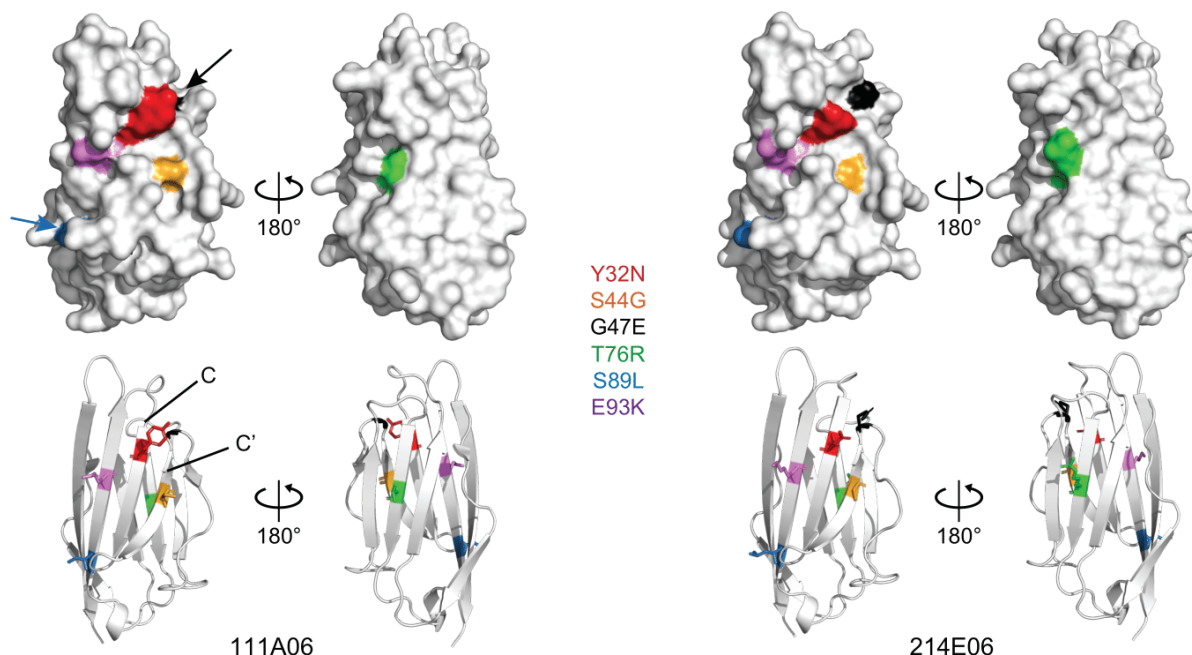


**Figure 53. pLDDT scores of predicted domain 1 structures.**

The 111A06 and Hap074 are shown as representative models. Color scale matches Table 11 and ranges from 70 to 100. Residues with pLDDT >90 are considered highly accurate and have their side chains oriented correctly 80% of the time. Residues with pLDDT >70 generally have their backbones predicted correctly.

The variable pLDDT score in G47/E47 (Table 11 and Figure 53) may be influenced by its position in the structure. It appears in the loop between the C' and C'' strands which may have more flexibility (Figure 54). Five mutations mapped to one face of the predicted  $\beta$ -sandwich (in strands C, C', and F) with the three specificity-altering mutations in close proximity to each other in  $\beta$ -strands C (N32/Y32) and C' (S44/G44) and the loop between C' and C'' (G47/E47) (Figure 54). This suggested these strands are involved in homophilic binding between compatible domain

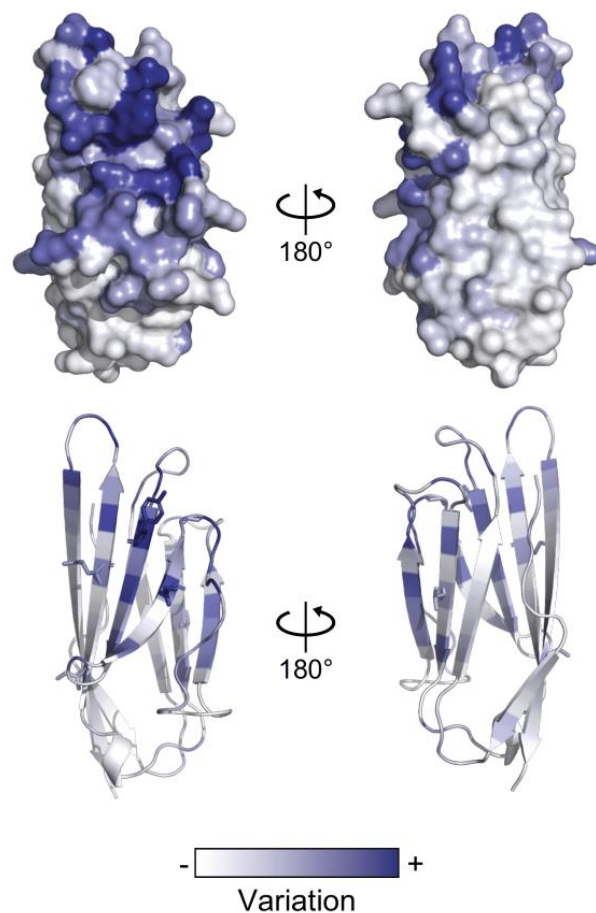
1 isoforms. The sixth variant residue (T76R) was present on the E strand on the opposite face of the domain.



**Figure 54. Predicted structure of Anc domain 1 with the six variant residues labeled in 111A06 and 214E06.**

Arrows point to only partially visible residues in the surface representation of 111A06. Strands C and C' and labeled in 111A06 for reference.

As one approach to identify functionally important parts in domain 1, I reasoned that selection should increase sequence variation at or near the binding site. I 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 111A06. I 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 55).

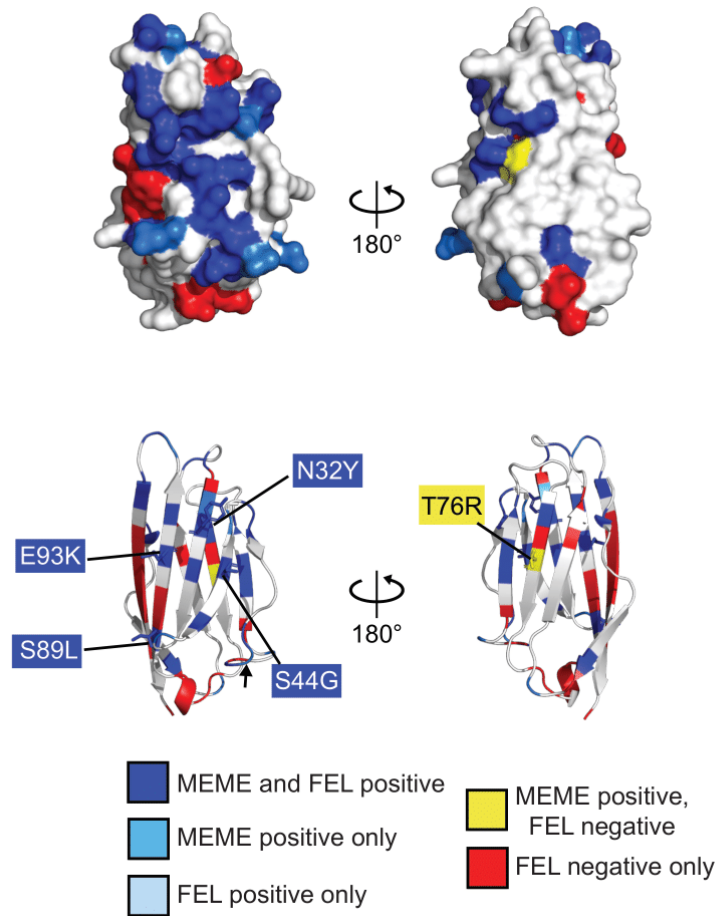


**Figure 55. Sequence conservation mapped onto domain 1.**

Conservation is mapped onto the 111A06 isoform as an example.

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), I was 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, I analyzed the alignment of all known domain 1 sequences with MEME (Murrell et al., 2012) and FEL (Pond & 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 56 and Table 12). Twenty sites were predicted to be under negative (purifying) selection and mapped to this side of the domain (Figure 56 and Table 12).



**Figure 56. Residues predicted to have experienced either diversifying or purifying selection mapped onto domain 1.**

The 111A06 isoform is shown as an example. Colors correspond to the predictions of MEME and/or FEL.

Arrowhead indicates the one residue predicted to be under positive selection by FEL only.



**Table 12. Sites in domain 1 under positive or negative selection.**

Codon in full alignment <sup>a</sup>	Corresponding codon in Anc <sup>a</sup>	HyPhy FEL (type of selection <sup>b</sup> , p-value)	HyPhy MEME (p-value)
3	3	Pos., 0.0069	0.0116
11	11	Pos., 0.0621	0.0830
12	12	Neg., 0.0585	
14	14		0.0000
21	21	Neg., 0.0001	
23	23	Neg., 0.0664	
35	26	Pos., 0.0510	0.0699
40	29	Pos., 0.0304	0.0014
42	30	Neg., 0.0126	
43	31		0.0014
44 <sup>c</sup>	32 (N32Y)	Pos., 0.0925	0.0131
46	34	Pos., 0.0427	0.0599
51	39	Pos., 0.0623	0.0831
53	41	Pos., 0.0039	0.0068
56 <sup>c</sup>	44 (S44G)	Pos., 0.0786	0.0000
58	46		0.0000
59 <sup>c</sup>	47 (G47E)		
60	48	Pos., 0.0191	0.0075
61	49	Pos., 0.0000	0.0000
68	55	Pos., 0.0041	0.0072
69	56	Pos., 0.0971	
70	57	Pos., 0.0387	0.0549
71	58	Neg., 0.0288	
73	60		0.0057
75	62	Neg., 0.0295	
77	64	Pos., 0.0676	0.0000
78	65	Pos., 0.0353	0.0507
80	67	Pos., 0.0018	0.0033
87	74	Pos., 0.0113	0.0182
88	75	Neg., 0.0064	
89 <sup>d</sup>	76 (T76R)	Neg., 0.0163	0.0005
95	82	Neg., 0.0021	
96	83	Neg., 0.0282	
98	85	Neg., 0.0005	
99	86	Neg., 0.1000	
102 <sup>d</sup>	89 (S89L)	Pos., 0.0410	0.0577
103	90	Neg., 0.0091	
104	91	Neg., 0.0005	
106 <sup>d</sup>	93 (E93K)	Pos., 0.0550	0.0525
109	95	Pos., 0.0174	0.0114

Table 12 continued on next page

**Table 12. Sites in domain 1 under positive or negative selection (continued)**

121	98		0.0000
122	99	Pos., 0.0132	0.0209
124	101	Pos., 0.0023	0.0042
125	102	Pos., 0.0269	0.0397
132	104	Neg., 0.0228	
133	105	Neg., 0.0263	
134	106	Neg., 0.0914	
135	107	Neg., 0.0092	
136	108	Neg., 0.0000	
138	110	Neg., 0.0049	
142	114	Neg., 0.0407	

<sup>a</sup> Gaps in the full alignment cause a difference in codon numbering between the full alignment and Anc.

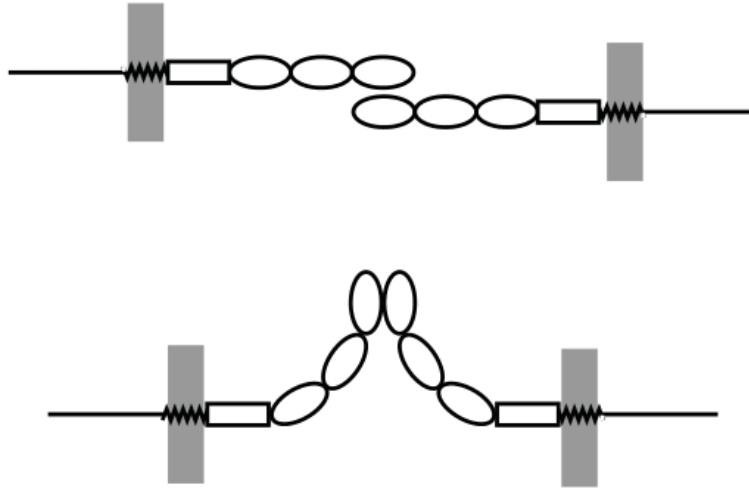
<sup>b</sup> Pos. = positive selection, Neg. = negative selection

<sup>c</sup> Site that determines specificity in this study

<sup>d</sup> Site that does not determine specificity in this study

With respect to the six positions at which mutations occurred in the 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. I speculate these interactions could occur in either an antiparallel or parallel topology (Figure 57).



**Figure 57. Hypothetical binding topologies Alr2.**

## 4.5 Discussion

Domain 1 is the most polymorphic region of Alr2 (Gloria-Soria et al., 2012). Here, I 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, I 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 confidence in the 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, the results show that natural selection maintains ancestral sequences alongside

one encoding new specificities. Thus, the results reveal a mechanism capable of generating, maintaining, and increasing the functional diversity of *Alr2*.

In this study, I 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<sup>a/null</sup>* (where *a* is an allele encoding a homophilic binding protein and null cannot bind homophilically) might be functionally equivalent to *Alr2<sup>a/a</sup>* 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 non-self is reduced from the sum of two allele frequencies to the frequency of a single allele. So why were null alleles not detected in this and a previous study (Karadge et al., 2015)? One possibility is that null alleles are rare, and I have not found one yet because I have only studied ~5% of sequence variation at *Alr2*. A second possibility is that *Alr2<sup>a/null</sup>* animals are not, in fact, equivalent to *Alr2<sup>a/a</sup>* animals. This might be true if *Alr2* has essential functions beyond self-recognition at the colony border. In fact, I 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<sup>wt/null</sup>* colonies might be unfit, and *Alr2<sup>null/null</sup>* 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 the assay is unable to detect null alleles. This would be the case if cells aggregate in the assay at a lower affinity than that required for colonies to recognize a tissue as self.

Assuming the 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 non-allorecognition genes tightly linked to *Alr2*, some *Alr2* alleles might have higher than expected frequencies due to genetic hitchhiking. In addition, I note that the tree for this clade does not represent actual allele frequencies because I 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.

The main limitation of this study is the qualitative nature of the 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 & 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 I 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 the 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, I conservatively interpreted “semi-mixed” aggregates as indicating that two isoforms had heterophilic affinities that were relatively weaker than their homophilic affinities. I 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 the 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).

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, the evolutionary analyses cannot tell us whether the six specific mutations that occurred within the branches of the clade experienced positive selection. What I 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 the assays could not detect. With respect to position 47, which was not found to be under positive selection but 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 the 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 to 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.

The exquisite specificity displayed by these closely related isoforms can be useful in modeling potential binding mechanisms. While AlphaFold generated high-quality structures, there were some areas of lower confidence, specifically in the C' and C'' strands, which suggested that although the backbone is likely in the proper orientation, the side chains may not be. To fully model and understand the biophysical mechanism of these domains, it must await experimentally determined structures. I was, 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 V-set  $\beta$ -barrel, primarily in C, C', and C'' strands. This suggests that the N-terminal domains of Alr2 bind in a side-to-side manner.

Although I 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 these results, *Hydractinia* would be a productive system in which to study protein epistasis or “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 & Kondrashov, 2010; Podgornaia & Laub, 2015). Because natural selection drives the continued evolution of new allorecognition alleles, allorecognition loci like Alr2 are essentially natural experiments exploring sequence space.

## **4.6 Methods**

### **4.6.1 Experimental model and subject details**

HEK293T cells (ATCC Cat# CRL-3216) were cultured at 37°C with 5% CO<sub>2</sub> 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).

#### **4.6.2 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, I 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](http://www.hiv.lanl.gov)), to yield 146 distinct domain 1 cDNA sequences, encoding 137 distinct amino acid sequences.

#### **4.6.3 Phylogenetic Analysis and Ancestral State Reconstruction**

The 146 domain 1 cDNA sequences were aligned with PRANK (Löytynoja, 2014), a codon-aware alignment program. 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). 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 (Dutheil & Boussau, 2008; Löytynoja, 2014). An unrooted tree was generated using iTOL v5.5.1 with one iteration of equal-daylight (Letunic & Bork, 2019).

#### **4.6.4 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 214E06 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  $\geq 20$  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.

#### **4.6.5 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  $3 \times 10^5$ /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  $\mu$ l using 1  $\mu$ g (X  $\mu$ l) of plasmid DNA (plasmid concentrations between 300ng-1000ng/ $\mu$ l), diluted with optiMEM (Gibco, #31985-070) (97-X  $\mu$ l), and 3  $\mu$ l of TransIT-293 reagent. While incubating the DNA:lipid complexes, the cells were washed using 500  $\mu$ l of DPBS (Fisher Scientific, BW17-512F), incubated with 1 ml transfection medium (complete HEK medium without antibiotics), and replaced in the 5% CO<sub>2</sub> incubator. Once the DNA:lipid complexes had incubated, the 100  $\mu$ l 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.

#### **4.6.6 Aggregation assay**

The 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 35 $\mu$ m strainer mesh (Stellar scientific, FSC-FLTCP)

to create a single cell suspension. For each aggregation assay, a total of  $5 \times 10^4$  cells were resuspended in 500  $\mu$ l 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,  $2.5 \times 10^4$  cells of each transfection were added to the same well and resuspended in a total of 500  $\mu$ l. The plate was incubated for 1 h at 37°C in 5% CO<sub>2</sub> 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  $\mu$ l 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).

#### **4.6.7 Sequence Variability and visualization of Domain 1**

The structure for the Alr2 domain 1 isoforms were predicted using AlphaFold (Jumper et al., 2021), as implemented in Colabfold (Mirdita et al., 2021) as before (3.6.3). Each *Alr2* sequence was added to the HHblits MSA query results obtained for Alr2 domain 1 (3.6.3). The top structural model of each prediction was submitted to the DALI server (<http://ekhidna2.biocenter.helsinki.fi/dali/>, Holm, 2020) to identify structures from the PDB (rcsb.org; Berman et al., 2000) with the same fold. The top hit was downloaded and if necessary modified to contain only the relevant chain for structural alignment. Each prediction was then submitted with its identified PDB structure to TMalign (<https://zhanggroup.org/TM-align/>, Zhang & Skolnick, 2005) to obtain the TM-score of the alignment. Secondary structure was predicted

with STRIDE (Heinig & Frishman, 2004). Structural models were visualized in PyMOL 2.0 (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.).

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; E. C. Meng et al., 2006). The conservation attributes were uploaded to the domain PDB file and rendered onto the Alr2 D1 structure in PyMOL. Sites under positive selection were identified using MEME (Murrell et al., 2012) and FEL (Pond & 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).

## **5.0 Extreme variation in gene sequence and copy number in the *Alr* gene family**

### **5.1 Foreword**

The work in this chapter is adapted from a manuscript in preparation for publication in which I am second author: Steven M. Sanders, Aidan L. Huene, Zhiwei Ma, Anh-Dao Nguyen, Sergei Koren, Adam Phillippy, Christine E. Schnitzler, Andreas D. Baxevanis, and Matthew L. Nicotra (2022, Unpublished)

### **5.2 Summary**

Following the annotation of the ARC-F haplotype, it was unknown how greatly the genomic architecture, gene content, and sequence variability of the ARC differed in other haplotypes, especially in non-inbred animals. To address these questions, I assembled, aligned, and annotated two ARC haplotypes from an outbred colony, 291-10. In this chapter, I show that the genomic architecture of the ARC appears to be largely the same between the F and one of the 291-10 haplotypes. The 291-10 ARC contains 54 unique *Alr* loci. Many *Alr* loci can be found in all three haplotypes. There was some variation in copy number and in gene classification. There was also high allelic polymorphism in *Alr1*, *Alr2*, and *Alr6*, while the remaining *Alr* loci were more conserved.

### 5.3 Introduction

The ARC-F sequence generated in Chapter 2.0 provided a wealth of information regarding the complexity and overall organization of the *Alr* family within the ARC. The sequence diversity between genes along with their conserved domain architecture suggests that there might be frequent duplication and recombination events within the ARC or high mutation rates that generated this diverse family of genes. Genes within the clusters appear to be duplicated more often than genes outside the clusters.

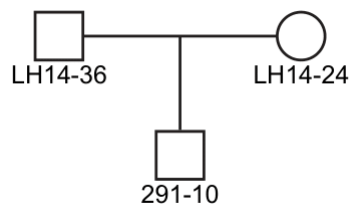
Several questions remain outstanding following the assembly and annotation of ARC-F. First, what is the full sequence of the ARC? There were two gaps of unknown size between the Cluster B and Cluster C regions and the *Alr37* and *Alr38* sequences were located on separate contigs whose linkage could not be determined based on the assembly. Second, how conserved is the genomic organization of the ARC? While generating the inbred lines, it was anticipated that the genetic background may become homogenized thus limiting the identification of any unlinked allorecognition loci. In addition, sampling only one ARC haplotype only provides one possible arrangement of allorecognition loci in that haplotype. Third, how static is the *Alr* content when comparing different haplotypes? Are the same genes localized to the same region? Do the *Alr* sequences identified change drastically between haplotypes? Are genes expressed differently based on their haplotype? While the conservation of *Alr1* and *Alr2* has been absolute in all tested lines to date, conservation of the other *Alr* sequences, particularly those which may not have a function such as pseudogenes, could not be predicted. Fourth, how much variation is present between alleles of the new *Alr* sequences? *Alr1* and *Alr2* have both been shown to have highly polymorphic alleles particularly within their ectodomain (Gloria-Soria et al., 2012; Karadge et al., 2015), but this may not be the case for all *Alr* sequences. To address these questions, I assembled

and annotated two ARC haplotypes from a heterozygous wildtype colony to compare it with the ARC-F.

## 5.4 Results

### 5.4.1 Assembly of the heterozygous 291-10 ARC reference sequence

To understand how the *Alr* content differs in other haplotypes, the genome of colony 291-10, an outbred male, was sequenced and assembled (Figure 58). High-molecular weight genomic DNA was sequenced at the NIH via PacBio long-read sequencing and high-throughput Illumina data. Sequences were assembled with Canu (Koren et al., 2017) and polished with the Illumina short read data using pilon (Walker et al., 2014). The resulting non-filtered assembly was 406 Mb long, with 4,480 scaffolds, an N50 of 2.2 Mb, and 22,022 predicted genes. These 4,480 scaffolds contained the full, diploid genome.



**Figure 58. Pedigree of the wildtype colony 291-10 used for genomic sequencing.**

Colony 291-10 is the offspring of two colonies, LH14-36 and LH14-24, collected from Lighthouse Point, New Haven, CT in 2014.

Unlike the laboratory-generated colony 236-21 which was homozygous for the ARC-F, 291-10 is heterozygous. To aid in identifying alternate haplotypes, the full, diploid assembly was



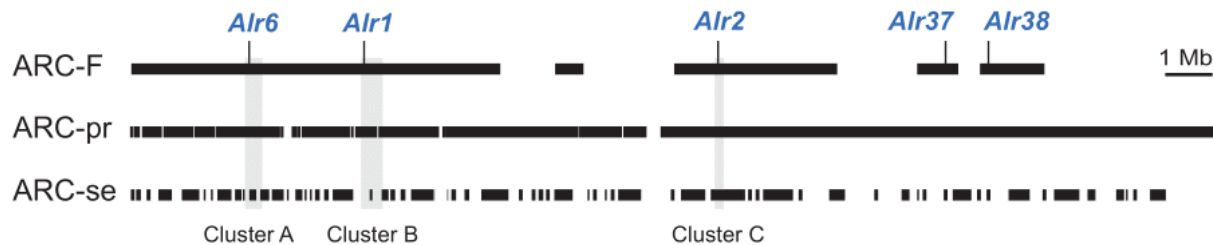
separated into “primary” and “secondary” assemblies. The primary assembly was filtered to include the largest contigs (~0.5-20 Mb) and theoretically represents the full haploid genome. This assembly was analyzed by BUSCO and was estimated to have only 11% duplicate genes. Therefore, some genes may have both alleles represented in the primary assembly. The secondary assembly contained the remaining contigs (usually <0.5 Mb) which either encoded the second allele of those sequences found in the primary assembly or sequences that could not be aligned.

Due to the heterozygous nature of the genome and the lack of parental genomes for 291-10, the contigs from these primary and secondary assemblies have not been phased. Thus, contigs from the primary assembly likely do not come from the same chromosomal set and likewise with contigs from the secondary assembly. The secondary assembly represents only those regions of the genome with enough variation that can be assembled into a separate contig. Therefore, the secondary assembly likely does not include all sequences which would represent a second haploid set of the genome. The regions of low variability which cannot be assembled as separate haplotypes would have been collapsed into the contigs that were filtered into the primary assembly. Knowing the correct genomic phasing is not necessary for studying the ARC sequence variability this chapter, however it may affect how many alleles can be recovered.

One difficulty with assembling any genome, particularly heterozygous genomes, is that complex regions containing repeats, numerous SNPs, and high levels of variation may be assembled into more than two haplotypes. The presence of more than two haplotypes can be due to misassembly but may in some cases represent large scale duplications which can be difficult to resolve depending on the sequence and genomic positioning data available.

To find all contigs that would align with the previously expanded ARC-F reference sequence, I used NUCmer to align the entire 291-10 assembly (both primary and secondary) to the

ARC-F sequence. I identified 16 contigs from the primary assembly (ARC-pr) and numerous secondary contigs (ARC-se) that aligned to the ARC-F reference (Figure 59). The primary assembly contigs were able to connect the gap between the Cluster A and B reference sequence and the short contig containing marker 28. In addition, one contig of sizeable length (11 Mb) was able to connect the ARC-F contigs encoding Cluster C, *Alr37*, and *Alr38*, showing that all *Alr* found in the ARC-F are linked in 291-10 and, likely, also in the ARC-F haplotype. In the contigs of the primary assembly, no large genomic rearrangements were observed (e.g., one contig from ARC-pr aligning to spatially separate regions in ARC-F). The contigs of the secondary assembly did not cover enough sequence to assess whether any genomic rearrangements may be present.



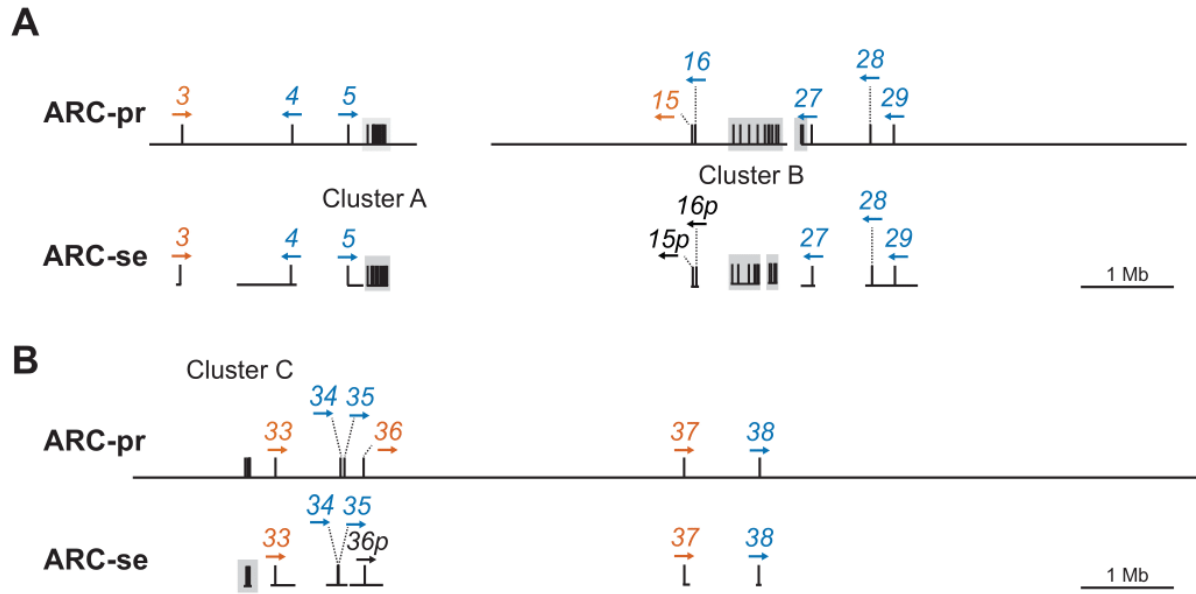
**Figure 59. ARC contig alignment between F reference sequence and the wildtype haplotypes.**

One *Alr* sequence from each cluster or contig in ARC-F is labeled for reference. Many ARC-se contigs are too short to be added to the alignment at this scale. All contigs bearing *Alrs* are represented in the diagram.

#### 5.4.2 The 291-10 ARC contains a larger set of the *Alr* family

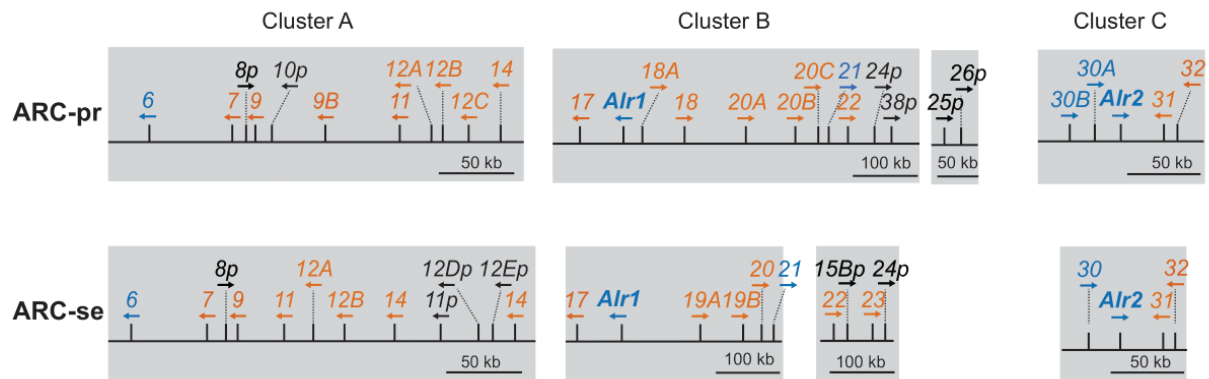
Next, I annotated all *Alr*-like sequences in the 291-10 assembly. Briefly, regions predicted to encode *Alr*-like sequences were identified using AUGUSTUS predictions (Stanke et al., 2004), mapped RNAseq data from colony 291-10, and BLASTX using an updated database including all *Alr*-like sequences annotated in the ARC-F haplotype. These data were loaded into the Apollo

annotation platform (Dunn et al., 2019) where I created gene models. To identify *Alr* genes that might exist outside the 291-10 contigs aligned to the ARC-F reference sequence, I used TBLASTX to query the full 291-10 genome assembly with the amino acid translation of each *Alr* gene model. No new contigs were identified beyond those that aligned to the ARC-F. *Alr* models were named according to their genomic location and sequence similarity to genes from the ARC-F haplotype. As observed in the ARC-F *Alr* gene models, the 291-10 *Alr* gene models varied in the amount of RNAseq support. For consistency, I classified the newly annotated *Alrs* as either a bona fide gene, a putative gene, or a pseudogene based on the previous criteria (0). To distinguish which assembly the *Alr* sequences come from, each *Alr* will have the assembly abbreviation connected to the end of its name with a hyphen (e.g., *Alr1-pr*, *Alr12A-se*). For those *Alrs* that had a third allele annotated, the assembly designation will be “*se2*”, (e.g., *Alr5p-se2*). A total of 90 *Alr* sequences, representing 54 unique *Alr* loci, were annotated on 25 contigs (Figure 60, Figure 61, and Table 13). Four contigs from the primary assembly contain about half (46) of the total *Alr* sequence annotations (Supplemental Table 1). Sixteen contigs, three from the primary assembly and thirteen from the secondary assembly, contain 38 *Alr* annotations which in most cases represents the alternate alleles of those genes identified in the primary assembly contigs (Supplemental Table 1). The remaining six *Alr* annotations were identified as a third allele (*Alr-se2*) and were present on one primary and four secondary assembly contigs (Supplemental Table 1). One sequence (*Alr1-se2*) was present on a contig which only included itself and was identical to the sequence of *Alr1-se*. As it represents a redundant sequence, it will be excluded from further analyses. The other five sequences appeared on contigs which appeared to be a misassembly or a chimeric assembly from the haplotypes present in the *ARC-pr* and *ARC-se*. Thirty-one *Alr* sequences were present in two haplotypes. Twenty-three annotations were only able to be annotated in one haplotype (Table 13).



**Figure 60. *Alr* annotations identified in 291-10.**

A) *Alr* annotations on contigs encoding Cluster A and Cluster B in both haplotypes. The four contigs from the primary assembly encoding B) *Alr* annotations on contigs encoding Cluster C.



**Figure 61. *Alr* annotations in Clusters.**

**Table 13. Alr annotations in the ARC.**

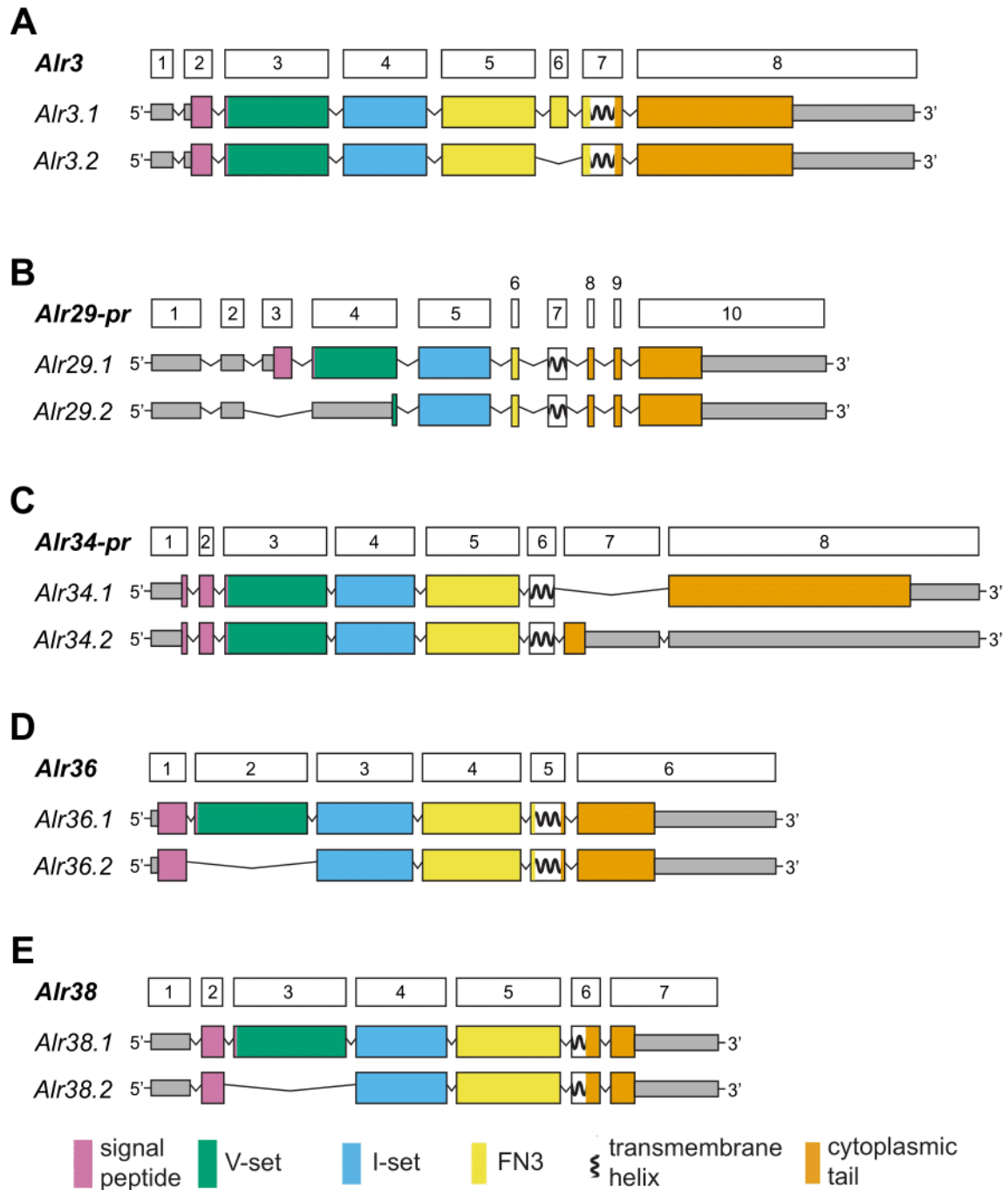
ARC-pr contains those annotations from the four primary contigs with contain 46 Alr annotations. ARC-se contains contigs from three primary contigs and thirteen secondary contigs which contain 37 Alr annotations. ARC-se2 contains any Alr annotations for which I found a third sequence.

ARC-pr	ARC-se	ARC-se2	ARC-pr	ARC-se	ARC-se2
Alr1	Alr1	Alr1	Alr20	Alr20p	
Alr2	Alr2		Alr20A		
Alr3	Alr3		Alr20B		
Alr4	Alr4		Alr20C		
Alr5p	Alr5p	Alr5p	Alr21	Alr21	
Alr6	Alr6		Alr22	Alr22	
Alr7	Alr7			Alr23	
Alr8p	Alr8p		Alr24p	Alr24p	
Alr9	Alr9		Alr25p		
Alr9B			Alr26p		
Alr10p			Alr27	Alr27	
Alr11	Alr11	Alr11	Alr28	Alr28	
	Alr11p		Alr29	Alr29	
		Alr12		Alr30	
Alr12A	Alr12A	Alr12A	Alr30A		
Alr12B	Alr12B		Alr30B		
Alr12C			Alr31	Alr31	
Alr12Dp			Alr32	Alr32	
Alr12Ep			Alr33	Alr33	
Alr14	Alr14	Alr14	Alr34	Alr34	
Alr15	Alr15p		Alr35	Alr35	
	Alr15Bp		Alr36	Alr36p	
Alr16	Alr16p		Alr37	Alr37	
Alr17	Alr17		Alr38	Alr38	
Alr18			Alr38p		
Alr18A					
Alr19					
	Alr19A				
	Alr19B				

There were several gene models whose full-length predicted amino acid sequences had >80% sequence identity. *Alr9*, *Alr12*, *Alr18*, *Alr19*, *Alr20*, *Alr15*, and *Alr38* all had at least two copies. For those genes present within a cluster, the duplicate genes were also present in the same

cluster. The copies of the two sequences present outside of a cluster, *Alr15* and *Alr38*, were translocated to within Cluster B (Figure 61).

Interestingly, I identified the same splice variants in *Alr1*, *Alr2*, and *Alr6* as I had identified previously in ARC-F (Figure 17 and Figure 18). This suggested that the splicing of these genes may be well conserved across haplotypes and spliced products likely have specific functions in *Hydractinia*. Several additional *Alr* genes exhibited alternative splicing or splice junction variants. Five *Alr* genes in one or both alleles had RNA-seq supporting the alternative splicing of an entire exon (Figure 62). Some of the most interesting splice variants among these are *Alr36* and *Alr38* which have alternative splice variants in both alleles that exclude the V-set domain (exon 2) but still have the signal peptide (exon 1) spliced to the I-set domain (exon 4) that could produce a functional protein (Figure 62D,E). The function of the I-set domain in an Alr has not yet been determined. In addition, the function of those domains that sit in such close proximity to the cell surface is not yet known.

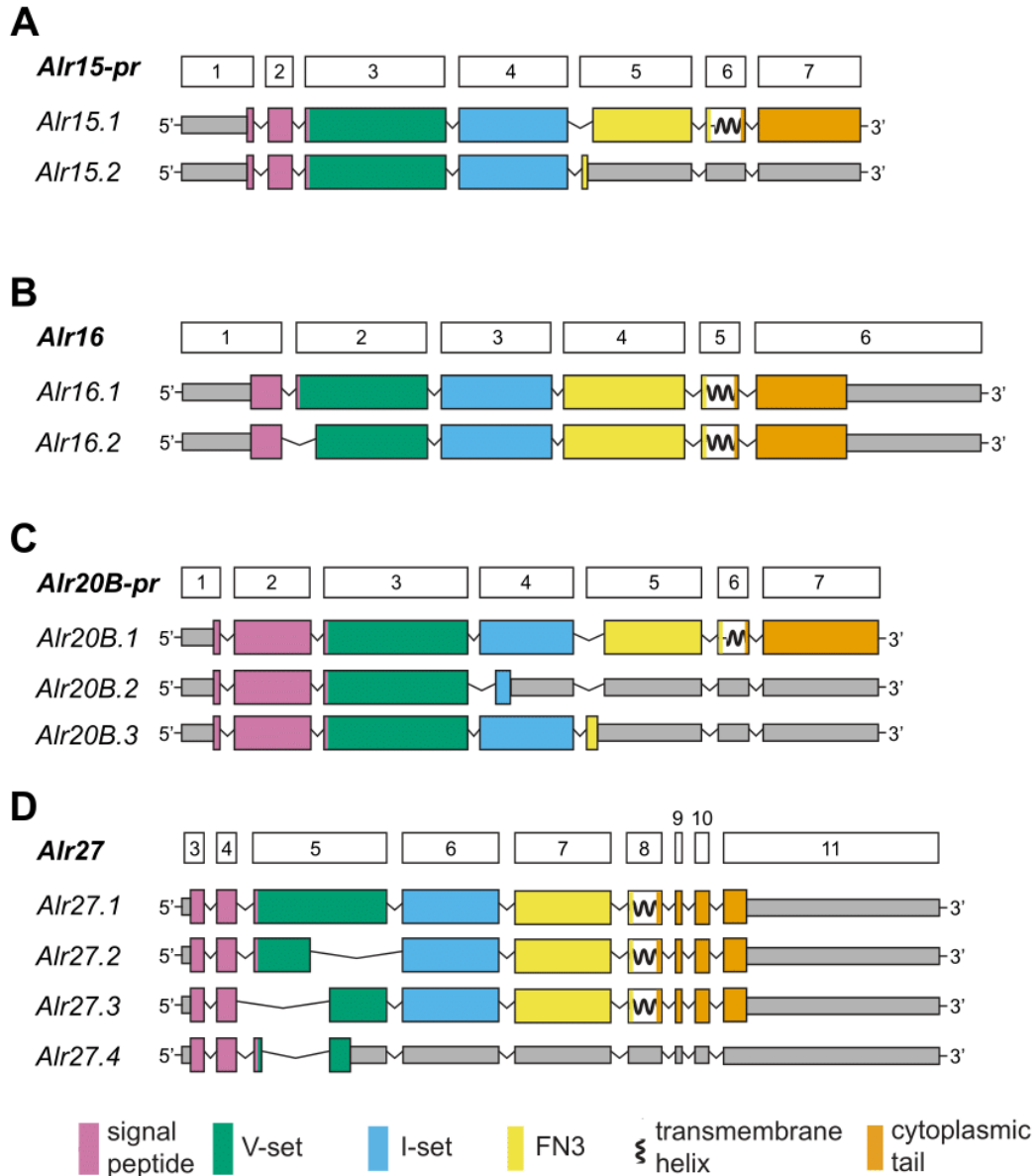


**Figure 62. Alternative splice variants involving entire exons.**

A) *Alr3* (*pr* and *se*) both have an additional splice variant which excludes exon 6 encoding an extended portion of the FN3 domain. B) *Alr29-pr* encodes a splice variant which would exclude exon 3 (encoding the signal peptide). C) *Alr34-pr* encodes a splice variant which excludes exon 7 resulting in a truncated cytoplasmic tail. D) *Alr36* (*pr* and *se*) and E) *Alr38* (*pr* and *se*) both encode splice variants which exclude exon 3 encoding the V-set domain.

In four *Alrs*, I found alternate splice junction patterns at one of their exons (Figure 63). The splice junction variants *Alr15.2-pr*, *Alr20B.2*, *Alr20B.3*, and *Alr27.4* resulted in premature stop codons and would only be expressed as secreted proteins. The splice variants of *Alr16.2*, *Alr27.2*, and *Alr27.3* did not affect the coding frame and would be expressed as transmembrane proteins. These splice junction variants may have a biologically relevant function, but this would require further testing to verify.

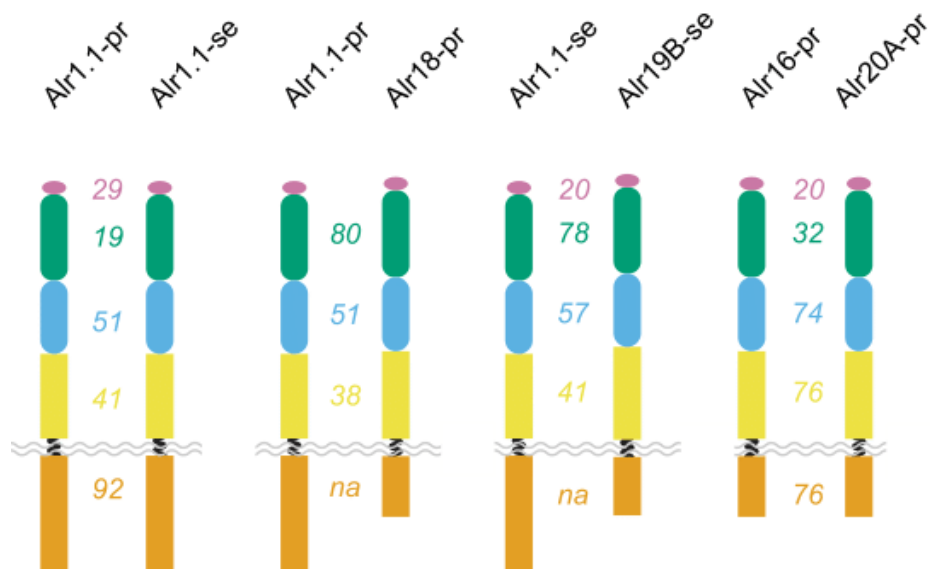




**Figure 63. *Alrs* with splice junction variants.**

A) Alr15-pr had a splice junction variant in exon 5. B) Alr16-pr and Alr16-se both have the same splice junction variant that occurs near the beginning of exon 2 which does not change the coding frame. C) Alr20B-pr has two additional splice junction variants which occur in exon 4 and 5 and results in a truncated sequence that would result in secreted isoforms. D) Alr27-pr has four splice junction variants. Alr27.1 encodes the full length protein, the remaining three have splice junction variants all within exon 5. Alr27.2 and Alr27.3 encode separate parts of exon 5. Only the splice junction variant in Alr27.4 results in a premature stop codon. Exon 1 (~2 kb) and exon 2 (<100 bp) only encode the 5' UTR and were not included in the graphic for simplicity.

To investigate the evolutionary relationships between *Alr* family members in ARC-pr and ARC-se, I analyzed the pairwise alignments of individual domains. I observed evidence for exon shuffling between several *Alr* gene pairs (Figure 64). *Alr1.1-pr* and *Alr1.1-se* encoded highly divergent sequences in all domains. Interestingly the V-set domain in *Alr1.1-pr* was fairly well conserved in *Alr18-pr* (80% sequence identity) whereas the *Alr1.1-se* was highly similar to the V-set domain in *Alr19B-se* (78% sequence identity) (Figure 64). These observations again support the presence of exon shuffling between *Alr1* and nearby *Alr*-like sequences (Rosa et al., 2010) as well as other *Alr* genes. The prevalence of exon shuffling in these haplotypes supports the hypothesis that exon shuffling is a common feature of *Alr* genes.



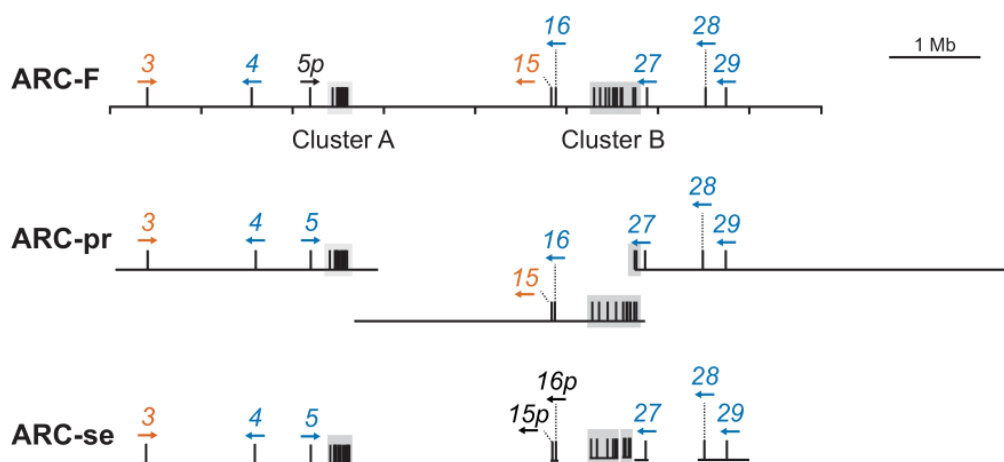
**Figure 64. Evidence of exon shuffling in *Alr* sequences.**

Numbers represent amino acid identity between domains. “na” represents sequences that were not alignable.

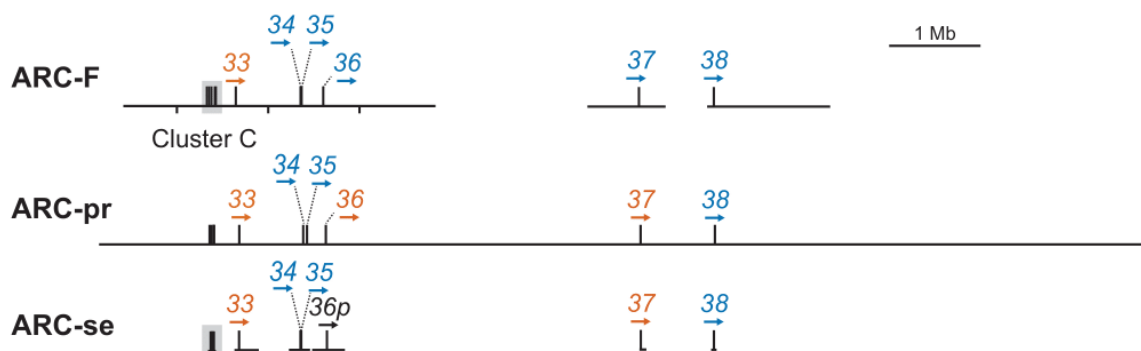
### 5.4.3 *Alr* content variation between 291-10 and ARC-F

To understand the *Alr* content variation between haplotypes, I aligned and compared all the annotations between ARC-pr, ARC-se, and ARC-F (Figure 65, Figure 66, and Supplemental Table 2). Of all of the annotations, only *Alr13p-F* was not identified in ARC-pr or ARC-se. No novel *Alr* sequences were identified beyond *Alr1-38*. Despite the lack of novel *Alr* sequences, the increased number of unique loci in ARC-pr and ARC-se (54) versus ARC-F (41) was one clear discrepancy. One reason for this difference in total loci was the presence of several duplicated genes present in the Clusters that were not present in the ARC-F (e.g. *Alr20A-pr* and *Alr20B-pr*) (Supplemental Table 2).

**A**

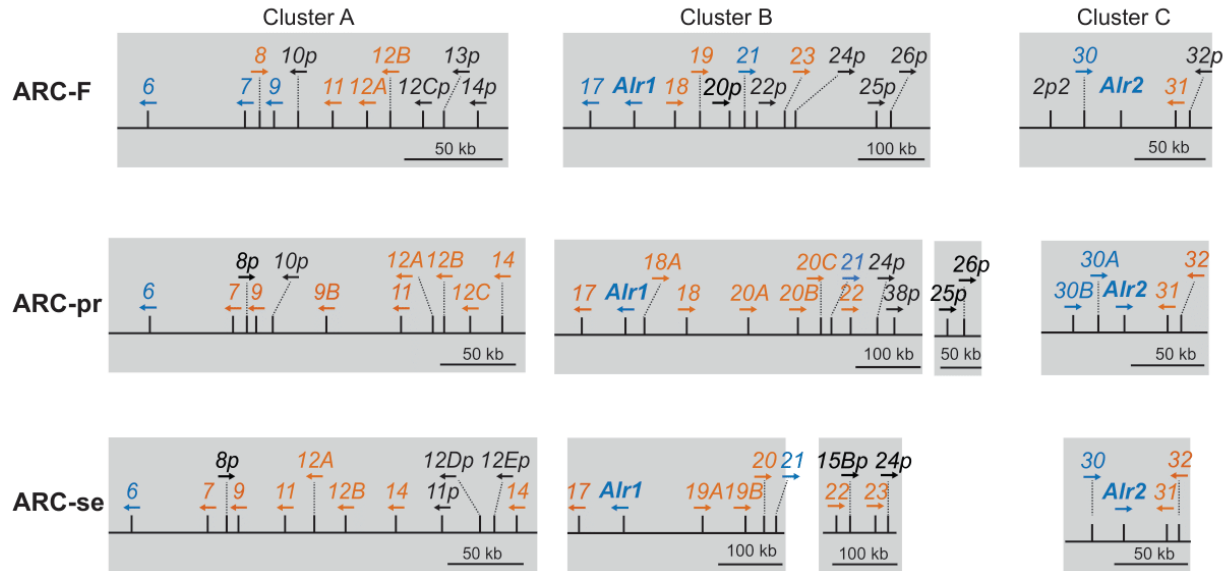


**B**



**Figure 65. Alignment between ARC-F, ARC-pr, and ARC-se.**

ARC-F, ARC-pr, ARC-se. A) Alignment between contigs encompassing Cluster A and Cluster B. B) Alignment between contigs encompassing Cluster C. The ARC-pr contig shows that Alr37 and Alr38 are physically linked with the ARC.



**Figure 66. ARC Clusters A, B, C alignment between F, pr, and se haplotypes.**

To systematically compare all the *Alr* sequences, I separated them based on the gene classification of the *Alrs* from ARC-F. Of the 18 gene models classified as bona fide in the *F*, 12 of them share the same bona fide classification in both 291-10 alleles (Table 14). Four of the genes lacked the RNA-seq coverage in both *pr* and *se* to classify them as bona fide genes and thus were classified as putative. *Alr16-pr* was classified as a bona fide gene. For *Alr16p-se*, the RNA-seq covered the entire coding sequence with evidence for proper splicing between exons 3-6. However, numerous reads aligned within the first two introns, only one spliced read was identified between exon 2 and 3, and two spliced reads were present between exon 1 and in the intron upstream of exon 2 that would result in a premature stop codon. Although this allele has been classified as a pseudogene, *Alr16p-se*, if properly spliced, would have been a bona fide gene. Moreover, all genes shared the same domain and exon-intron architecture between haplotypes. To capture all possible sequence variation present between these alleles, the properly spliced coding sequence of *Alr16p-*

*se* will be included later in sequence-level variation analyses. In *Alr36-pr*, all exons were fully covered by RNA-seq, but had an additional splice variant compared to *Alr36-F* (Figure 62). *Alr36-pr* had numerous intronic reads in the first intron and no evidence of splicing between exon 1 and 2 but had numerous spliced reads that supported splicing between exon 1 and 3. Although the predicted product of *Alr36-pr* is different than that of *Alr36-F*, *Alr36-pr* was classified as a bona fide gene. *Alr36p-se* had some RNA-seq coverage but was classified as such due to a 2-bp deletion at the end of exon 3 that resulted in a premature stop codon.

**Table 14. *Alr* bona fide gene classification comparison between the F, pr, and se haplotypes.**

Classification is based on the coloring scheme as before (bona fide in blue, putative in orange, pseudogenes in black).

<i>F</i>	<i>pr</i>	<i>se</i>
Alr1	Alr1	Alr1
Alr2	Alr2	Alr2
Alr4	Alr4	Alr4
Alr6	Alr6	Alr6
Alr7	Alr7	Alr7
Alr9	Alr9	Alr9
Alr16	Alr16	Alr16p
Alr17	Alr17	Alr17
Alr21	Alr21	Alr21
Alr27	Alr27	Alr27
Alr28	Alr28	Alr28
Alr29	Alr29	Alr29
Alr30		Alr30
Alr34	Alr34	Alr34
Alr35	Alr35	Alr35
Alr36	Alr36	Alr36p
Alr37	Alr37	Alr37
Alr38	Alr38	Alr38

Ten of the twelve models classified as putative genes in ARC-F are also putative in the ARC-pr and ARC-se assemblies (Table 15). Two of these genes (*Alr18-pr* and *Alr19-pr*) were

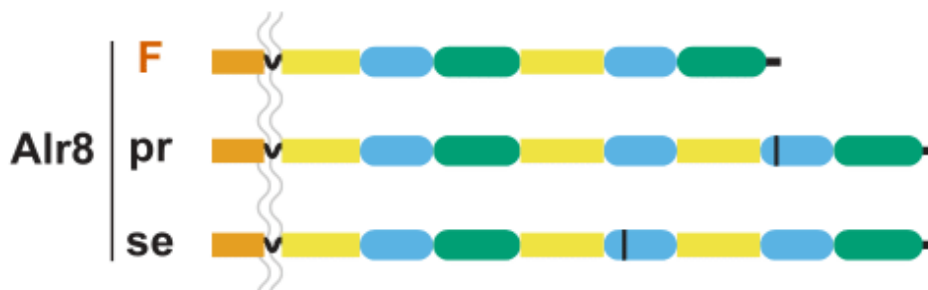
only present in one haplotype. *Alr19-pr* was present on a contig which only included its coding sequence, thus the localization of *Alr19-pr* is unknown currently. *Alr15p-se* was classified as a pseudogene due to a 1-bp deletion near the end of exon 3, which resulted in a premature stop codon. *Alr8p-pr* has a 1-bp deletion in exon 3 causing a premature stop codon. *Alr8p-se* has a 2-bp deletion in exon 5, also causing a premature stop codon. In addition, both *Alr8p-pr* and *Alr8p-se* possessed a variant domain architecture from *Alr8-F* (Figure 67).

**Table 15. *Alr* putative gene classification comparison between the F, pr, and se haplotypes.**

Classification is based on the coloring scheme as before (bona fide in blue, putative in orange, pseudogenes in black).

<i>F</i>	<i>pr</i>	<i>se</i>
<i>Alr3</i>	<i>Alr3</i>	<i>Alr3</i>
<i>Alr8</i>	<i>Alr8p</i>	<i>Alr8p</i>
<i>Alr11</i>	<i>Alr11</i>	<i>Alr11</i>
<i>Alr12A</i>	<i>Alr12A</i>	<i>Alr12A</i>
<i>Alr12B</i>	<i>Alr12B</i>	<i>Alr12B</i>
<i>Alr15</i>	<i>Alr15</i>	<i>Alr15p</i>
<i>Alr18</i>	<i>Alr18</i>	
<i>Alr19</i>	<i>Alr19</i> <sup>a</sup>	
<i>Alr23</i>	<i>Alr23</i>	<i>Alr23</i>
<i>Alr31</i>	<i>Alr31</i>	<i>Alr31</i>
<i>Alr33</i>	<i>Alr33</i>	<i>Alr33</i>
<i>Alr37</i>	<i>Alr37</i>	<i>Alr37</i>

<sup>a</sup> The location of *Alr19-pr* is unknown. It was present on a contig that only contained its sequence.



**Figure 67. Variant Alr8 domain architecture.**

Alr8p-pr and Alr8p-se both have an additional I-set-FN3 domain pair inserted between the first and second full ectodomains. Black line indicates the position of the deletion which causes premature stop codons.

Twelve gene models were present as pseudogenes in the ARC-F (Table 16). Of the 12 gene models classified as pseudogenes in the ARC-F, only *Alr13p-F* was not present in the ARC-pr and ARC-se assemblies. *Alr13p-F* is located in Cluster A and as such may be subject to higher rates of mutation or frequent rearrangements which could explain why it is absent in the ARC-pr and ARC-se assemblies. Five models were consistently present as pseudogenes. Two were able to be annotated in both haplotypes (*Alr24p* and *Alr5p*). *Alr10p-se* is located within Cluster A. Because there are two haplotypes for Cluster A which appear to be fully assembled, it is possible that it may have been deleted from the cluster in ARC-pr. Alternatively, if a *Alr10p-pr* were present in the cluster at one time, the sequence may have degraded over time to the point that it cannot be identified using the methods employed here. Both *Alr25p-pr* and *Alr26p-pr* were only identified in ARC-pr. Both are located at the end of Cluster B but are present on a separate contig from the rest of the cluster. While it is possible only one copy of each exists in the ARC, it is possible that there was not enough sequence diversity to generate two contigs for this region. The remaining seven models were present in one or both ARC-pr and ARC-se assemblies but were able to be



classified as bona fide or putative genes. The *Alr-se* sequence which corresponded in position and sequence to *Alr2p2-F* was fully expressed and spliced leading to its classification as a bona fide gene. To avoid nomenclature confusion with alleles of the *Alr2* gene, this *Alr* is named *Alr30B-se* (Table 16).

**Table 16. *Alr* pseudogene classification comparison between the F, pr, and se haplotypes.**

Classification is based on the coloring scheme as before (bona fide in blue and putative in orange).

<i>F</i>	<i>pr</i>	<i>se</i>
Alr2p2		<b>Alr30B<sup>a</sup></b>
Alr5p	Alr5p	Alr5p
Alr10p		Alr10p
Alr12Cp	<b>Alr12C</b>	
Alr13p		
Alr14p	<b>Alr14</b>	<b>Alr14</b>
Alr20p	<b>Alr20</b>	
Alr22p	<b>Alr22</b>	<b>Alr22</b>
Alr24p	Alr24p	Alr24p
Alr25p	Alr25p	
Alr26p	Alr26p	
Alr32p	<b>Alr32</b>	<b>Alr32</b>

<sup>a</sup> *Alr2p2-F* is the same sequence as *Alr30B-se*.

The ARC-pr and ARC-se sequences contained several additional sequences that did not match one of the loci from the ARC-F (Table 17). All of these had amino acid sequence similarity >80% to one of the known sequences. *Alr15* and *Alr38* were present outside the clusters in all three haplotypes but their duplications were translocated into Cluster B. Duplications of genes originally residing in the clusters remained in the same cluster. *Alr19A-se* contained a nearly identical duplicate domain 1 in its coding sequence which was not present in *Alr19-F* and *Alr19B-se* (Figure 68).

**Table 17. Additional *Alr* annotations found exclusively in the 291-10 haplotype.**

<i>pr</i>	<i>se</i>	Location relative to parent gene
	<b>Alr9B</b>	Within the same cluster
Alr11p		Within the same cluster
Alr12Dp		Within the same cluster
Alr12Ep		Within the same cluster
	Alr15Bp	Translocated into Cluster B
<b>Alr18A</b>		Within the same cluster
	<b>Alr19A</b>	Located in Cluster B, contains duplicated domain 1
	<b>Alr19B</b>	Located in Cluster B
<b>Alr20A</b>		Within the same cluster
<b>Alr20B</b>		Within the same cluster
<b>Alr20C</b>		Within the same cluster
Alr38p		Translocated at the end of Cluster B



**Figure 68. Alr19A-se has a duplicate domain 1 in its sequence.**

Interestingly, the ARC-F, ARC-pr, and ARC-se all contained the same splice variants for *Alr1*, *Alr2*, and *Alr6*. RNA-seq data for ARC-pr and ARC-se supported splice variants for several additional genes (Figure 62, Figure 63) that were only present as single transcript in the ARC-F. The discrepancy may have biological significance in the haplotypes compared here, however given that the RNA-seq data used for these analyses was to guide annotation and did not contain

biological or technical replicates, it is more likely that any splice variants not previously identified simply did not have the sequence coverage to be assembled.

#### **5.4.4 High allelic polymorphism at *Alr1*, *Alr2*, and *Alr6***

In previous studies, *Alr1* and *Alr2* have both been shown to be highly polymorphic (Gloria-Soria et al., 2012; Karadge et al., 2015). To determine whether any of the new *Alr* loci may also be highly polymorphic, I assessed their allelic variation between the ARC-F, ARC-pr, and ARC-se haplotypes. All full length sequences had some level of variation between the two or three haplotypes compared. Only in the case of four pairwise comparisons were the full length alleles identical: *Alr16-F* and *Alr16-pr*, *Alr22-pr* and *Alr22-se*, *Alr23-pr* and *Alr23-se*, *Alr27-F* and *Alr27-pr*. *Alr16* and *Alr27* are both outside of Cluster B. The presence of identical isoforms in two unrelated haplotypes suggests that these genes may be fairly well-conserved or have low sequence variation in most haplotypes. Similar to what has been observed studies with *Alr1* and *Alr2* (Karadge et al., 2015), most genes have higher levels of variation within the ectodomain than in their cytoplasmic tail. Therefore, I separated the ectodomains and cytoplasmic tails and compared the alleles pairwise (Table 18 and Table 19). In the ectodomain comparison (Table 18), aside from *Alr1* and *Alr2*, *Alr6* had the next highest number of polymorphic sites accounting for 34.3% of the sites. The gene with the fourth highest number of polymorphic sites was *Alr12B* with 19.2% sites being variable. Four genes had polymorphic sites accounting for 10-15% while the rest were under 10%. In the cytoplasmic tail (Table 19), only two genes had greater than 10% polymorphic sites. While the complete role of the cytoplasmic tail has not yet been fully studied in relation to allorecognition or to these proteins in general, the lower rates of polymorphism do suggest a potentially conserved function among *Alrs*.

**Table 18. Amino acid variation in the ectodomain.**

*Alr* genes are colored when all alleles compared are classified the same. *Alr* genes in black compare multiple gene classifications or are pseudogenes.

Gene	Polymorphic sites		Length	Alleles compared <sup>c</sup>
	(%) <sup>a</sup>	(#) <sup>b</sup>		
<i>Alr1</i>	62.7%	210	335-353	3
<i>Alr2</i>	44.8%	184	398-411	3
<i>Alr3</i>	2.4%	8	341	3
<i>Alr4</i>	1.5%	5	337	3
<i>Alr5</i>	0%	0	208	2
<i>Alr5p</i>	8.1%	10	123-168	2
<i>Alr6</i>	34.3%	108	314-315	3
<i>Alr7</i>	3.6%	12	330	3
<i>Alr9</i>	6.8%	22	326	3
<i>Alr11</i>	12.8%	42	327	3
<i>Alr12A</i>	12.8%	41	321	3
<i>Alr12B</i>	19.2%	62	311-323	3
<i>Alr15</i>	8.0%	25	304-314	2
<i>Alr16</i>	1.8%	6	342	3
<i>Alr17</i>	3.6%	5	139	3
<i>Alr18</i>	14.2%	47	332	2
<i>Alr21</i>	3.7%	13	353	3
<i>Alr22</i>	1.5%	5	324	3
<i>Alr23</i>	6.7%	23	344	3
<i>Alr27</i>	9.7%	30	309	3
<i>Alr28</i>	2.5%	8	317	3
<i>Alr29</i>	0.9%	2	222	3
<i>Alr30</i>	14.8%	60	404-406	2
<i>Alr31</i>	1.8%	6	333	3
<i>Alr32</i>	5.8%	15	255-257	3
<i>Alr33</i>	4.0%	13	325	3
<i>Alr34</i>	1.2%	4	325	3
<i>Alr35</i>	1.3%	4	318	3
<i>Alr36</i>	2.9%	13	451	3
<i>Alr37</i>	2.5%	5	203	3
<i>Alr38</i>	0.6%	2	356	3

<sup>a</sup> The shorter length is used for calculation. Genes with >25% polymorphic sites are highlighted in green.

<sup>b</sup> Gaps are not counted as polymorphic sites.

<sup>c</sup> Certain genes are not present in all three haplotypes. See Supplemental Table 2 for these genes.

**Table 19. Amino acid variation in the cytoplasmic tail.**

Gene	Polymorphic sites		Length	Alleles compared <sup>c</sup>
	(%) <sup>a</sup>	(#) <sup>b</sup>		
Alr1	9.0%	14	155	3
Alr2	8.6%	19	221	3
Alr3	3.4%	6	178	3
Alr4	1.8%	2	113	3
Alr5	2.7%	2	75	2
Alr6	4.2%	5	118	3
Alr7	1.4%	1	69-71	3
Alr9	9.6%	7	73	3
Alr11	8.2%	6	73	3
Alr12A	4.1%	3	73	3
Alr12B	5.5%	4	73	2
Alr14 <sup>d</sup>	15.5%	11	71	3
Alr15 <sup>e</sup>	0.0%	0	89	3
Alr16	0.0%	0	45-78	3
Alr17	0.0%	0	10	3
Alr18	12.4%	10	81	2
Alr19	1.3%	1	78	2
Alr21	2.5%	2	80	3
Alr22 <sup>f</sup>	7.4%	6	81	3
Alr23	7.8%	4	51	3
Alr27	0.0%	0	47	3
Alr28	0.9%	1	112	3
Alr29	1.8%	2	109	3
Alr30	2.4%	2	82	2
Alr31	0.9%	1	107	3
Alr32 <sup>g</sup>	6.1%	7	115	2
Alr33	7.3%	4	55	3
Alr34	3.4%	10	296	3
Alr36	1.3%	1	75	2
Alr37	0.0%	0	37	3
Alr38	6.3%	3	46-48	3

<sup>a</sup> The shorter length is used for calculation. Genes with >10% polymorphic sites are highlighted in green.

<sup>b</sup> Gaps are not included in as part of the number of polymorphic sites

<sup>c</sup> Certain genes are not present in all three haplotypes. See Supplemental Table 2 for these genes.

<sup>d</sup> Compares Alr14p-F with Alr14-wt1 and Alr14-wt2 cytoplasmic tails.

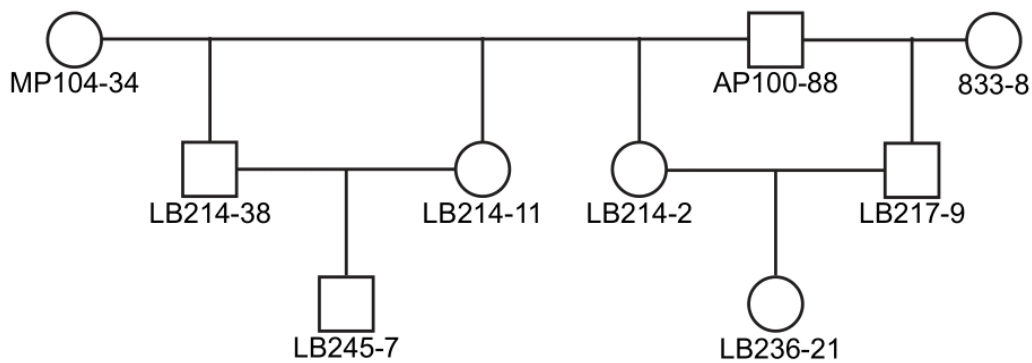
<sup>e</sup> Compared Alr15p-F with Alr15-wt1 and Alr15-wt2 cytoplasmic tails.

<sup>f</sup> Compares Alr22p-F with Alr22-wt1 and Alr22-wt2 cytoplasmic tails.

<sup>g</sup> Compares Alr32p-F with Alr32-wt1 and Alr32-wt2 cytoplasmic tails.

#### 5.4.5 The region including *Alr 3-14* was homogenized between haplotypes in inbred lines

After observing the high levels of sequence variation and copy number variation among the *Alrs* between the ARC-pr, ARC-se, and ARC-F, it was still unclear why any sequences from Cluster A had not been identified in previous screens for *Alr* candidates. To understand why this may have occurred, I obtained sequence from the genome of the second inbred line in which all previous work on allorecognition has been performed (Mokady & Buss, 1996; Cadavid et al., 2004; Powell et al., 2007; Nicotra et al., 2009; Rosa et al., 2010), 245-7 (Figure 69). This colony is homozygous for an alternative ARC haplotype, referred to as “ARC-R”.



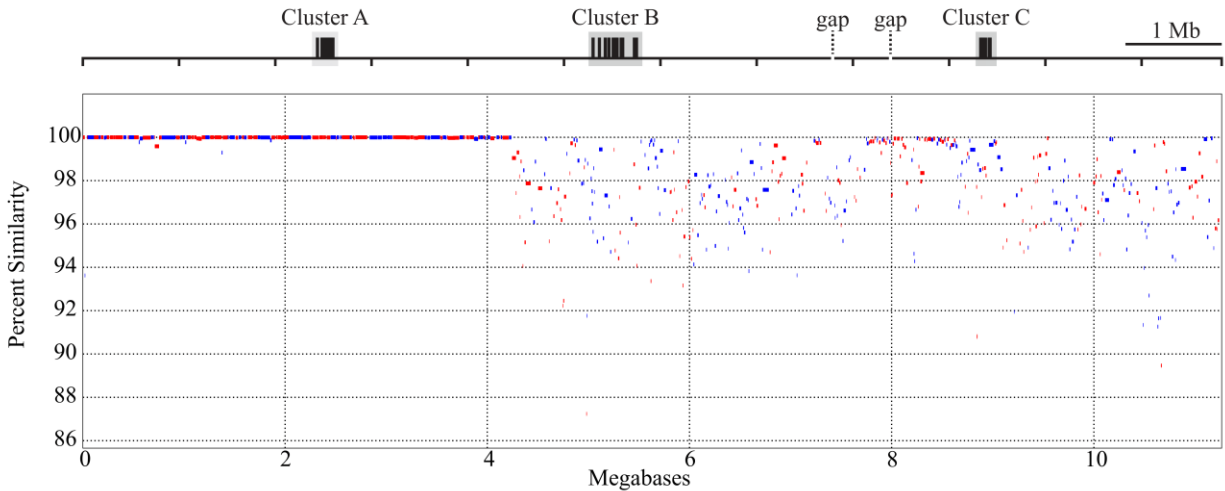
**Figure 69. Pedigree and relationship of the inbred colonies 245-7 (ARC-R homozygous) and 236-21 (ARC-F homozygous).**

Colony LB245-7 (ARC-R homozygous) shares one grandparent with LB236-21 (ARC-F genome animal). See

Figure 11 for full pedigree of parent colonies MP104-34, AP100-88, and 833-8.

The genome of colony LB245-7 was sequenced with Illumina paired-end sequencing, then assembled using DISCOVAR *de novo* (Weisenfeld et al., 2014). As expected, the resulting assembly was highly fragmented with 168,232 contigs. I used NUCmer to align these short contigs against the ARC-F reference sequence (Figure 70). ARC-R contigs that aligned to the Cluster A

region and the ~4.2 Mb surrounding it, including *Alr3-14*, were 100% identical to the ARC-F sequence. A small number of ARC-R contigs aligned with less than 100% identity, but these were repetitive regions that are likely misassembled in the short-read dataset (data not shown). These data are consistent with the ARC-F and ARC-R haplotypes having been homogenized during the generation of the inbred strains. The homogenization of *Alr3-14* between ARC-F and ARC-R is consistent with the hypothesis that an additional allodeterminant exists in this region, but was undetected in the inbred lines because it was not polymorphic.



**Figure 70. The region surrounding Cluster A was homogenized between ARC-R and ARC-F.**

Sequences aligned were a minimum of 1 kb and had 86-100% identity. Reads in red aligned with the forward strand and reads in blue aligned with the reverse strand.

## 5.5 Discussion

Previously, I determined the ARC-F reference sequence included at least 11.83 Mb of sequence that was comprised of five contigs, with two gaps of unknown size, and 41 *Alr* sequences,

of which I was unable to determine the linkage of *Alr37* and *Alr38* (2.0). Here, I annotated 54 unique *Alr* loci (90 annotations total) and show that nearly all *Alrs* previously identified in ARC-F are conserved in at least one ARC haplotype from the wildtype colony 291-10 and all *Alrs* are physically linked on the same chromosome within 18 Mb of sequence with only one gap of unknown size between Cluster B and Cluster C.

In comparing the *Alr* content between the 291-10 and ARC-F sequence, I found that nearly all genes were represented with the exception of *Alr13p-F*. The 291-10 genome did contain 12 additional annotations that were copy number variants of known *Alrs*. Observing each copy number variant in only one haplotype could be the result of the genomic assembly collapsing alternate alleles into the same contig due to lack of sequence variation or could represent a single duplication in only one haplotype. In either case, the presence of additional copy number variants within the ARC suggests that this region, particular the clusters, may be subjected to frequent duplication, deletions, and small-scale rearrangements that could impact the *Alr* diversity generated.

I used the same gene classification system as previously defined (2.0) as a baseline to compare the genes between haplotypes. Most of the bona fide and putative genes stayed within these two classifications suggesting that they may be conserved in other ARC haplotypes also. Interestingly, several of the previously identified pseudogenes in the F haplotype, usually lacking a predicted signal peptide or possessing a premature stop codon, were present as putative or bona fide genes in both wildtype haplotypes. Most of these sequences are present in the clusters suggesting that they may be subject to higher mutation rates or frequent rearrangements resulting in frequent pseudogenization. Alternatively, their presence as pseudogenes in the F haplotype may be an artifact of the inbreeding process. Sequencing the parental genomes is not possible because



they are no longer alive. However, sequencing additional wildtype ARC haplotypes may help elucidate how often these sequences are present as pseudogenes.

From studying the allelic variation in these three haplotypes, I found that in addition to *Alr1* and *Alr2*, *Alr6* also had a highly polymorphic ectodomain. Polymorphism is often associated with self/non-self recognition systems and may be a key trait that determines recognition specificity. This makes *Alr6* an attractive candidate for recognition specificity. In addition, *Alr6* is located within Cluster A. Several new *Alrs* (*Alr3-14*) are also present in this region which represent candidates for an additional allodeterminant. Many of the *Alr* alleles I compared had low allelic polymorphism. There are at least three reasons why this may be the case. First, I am only comparing the three haplotypes I have available for these genes. It is possible that additional haplotypes may show these *Alrs* can also be highly polymorphic. However, given the levels of polymorphism observed in the three alleles of *Alr1* and *Alr2*, it would be surprising to find a sudden increase in variation in additional alleles that was not observed in the three studied here. Second, low polymorphism could indicate that these genes are under weaker selection than *Alr1* and *Alr2*. Third, the lower levels of polymorphism could indicate a protein with a more conserved function. For example, both *Alr1* and *Alr2* encode proteins that have cell adhesion properties. While not yet determined, preliminary work with some of the new *Alr* genes suggests that they may also homophilically bind *in vitro*. An *Alr* with low or no polymorphism could function as a cell adhesion protein between all cells, and more interestingly, between different colonies, which would likely be necessary in order to create a strong connection.

The ARC-pr and ARC-F appear to be very similar in genomic organization as no large-scale rearrangements were obvious when comparing the aligned contigs. As the ARC-se contains more of the short contigs from the assembly, it is unclear whether genomic rearrangements are

present in at least one haplotype. Sequencing additional haplotypes, particularly those that are homozygous for the ARC, may help elucidate this. The similarities in *Alr* sequence identity between 291-10 and ARC-F and *Alr* localization patterns support the conclusions made previously that the *Alr* gene family may have arisen from duplication events that are relatively ancient and have potentially undergone substantial sequence evolution

In order to determine why the Cluster A and surrounding sequence was not identified when *Alr1* and *Alr2* were mapped (Cadavid et al., 2004), I sequenced genomic DNA from the 245-7 (R) inbred line for comparison. I demonstrated that the region surrounding Cluster A was homogenized between the F and R ARC haplotypes. The presence of a recombination breakpoint would explain how the region became homogenized since both 245-7 and 236-21 have shared ancestry (Figure 69). This result explains why if any *Alr*-like sequences or allodeterminants from Cluster A or the surrounding sequence were involved in allorecognition, they would not have been identified in a screen for phenotypic variants between these two colonies. If the predicted recombination point, or any other recombination points within the ARC exist, it could prove to be a useful tool for generating additional diversity for allorecognition. Future studies involving crosses with the known haplotypes across the ARC would aid in identifying additional recombination breakpoints.

## **5.6 Methods**

### **5.6.1 Sequencing and assembly of the genome of colony 291-10**

Colony 291-10, a wildtype male *Hydractinia* colony, was maintained as detailed in 2.6.1. DNA extraction was performed as detailed in 2.6.1. PacBio filtered subreads were generated with

the PacBio SMRTportal subread filtering protocol using default parameters. This process generated a single subread fastq file for each PacBio library sequenced. These filtered subreads were used as input to our genome assembly pipeline. The Canu (Koren et al., 2017) assembler was used to assemble the PacBio sequence data.

Canu assemblies were carried out using Canu v1.3 (<https://github.com/marbl/canu>) with default parameters. The program attempted to separate out contigs representing alternative haplotypes into primary and secondary assemblies via a filtering step. Due to the medium level of heterozygosity in both genomes, this filtering was not entirely successful, and the initial primary assemblies were larger than the expected haploid genome size with some contigs still representing duplicated loci from alternative alleles. The total assembly size for *H. symbiolongicarpus* was 731.169 Mb. The presence of duplicated loci in the initial primary assemblies was confirmed with BUSCO (Simao et al., 2015) v1.22, which indicated 42% duplicated genes. To remove much of the duplication and attempt to better separate haplotypes, self-alignments of all contigs with >1 read was performed with MUMmer 3.23 (Kurtz et al., 2004) with the command “nucmer –maxmatch –l 100 –c 1000 asm.ctg.fasta asm.ctg.fasta”. The number of matches > 5 kbp and 90% identity between all pairs of contigs was calculated and contig pairs were sorted by the number of matches. The contigs were greedily assigned to “primary” and “secondary” assemblies starting with the pair with the highest number of matches. Contigs with no alignments were then added to the secondary set. This generated a primary set of 395.756 Mbp and a secondary set of 335.412 Mbp. Following this filtering procedure, the presence of duplicated loci in the primary set according to BUSCO was reduced to 11%.

Scaffolding was done by Dovetail HiRise scaffolding with Illumina Chicago libraries constructed from the same gDNA extracted for PacBio and Illumina sequencing described above.

The primary set of contigs from Canu were sent to Dovetail. There were 5591 input contigs from the primary Canu set. After Dovetail scaffolding, there were 4611 scaffolds.

PBJelly software (<https://sourceforge.net/p/pb-jelly/wiki/Home/>) (English et al., 2012) from the PBSuite was used for gap filling the Canu-Dovetail assemblies using the PacBio reads. The program was run with gapInfo.bed files provided by Dovetail and the following parameters: -i --minGap=3. After the gap filling step, the assemblies had remarkably low percentages of remaining gaps: the *H. symbiolongicarpus* assembly had 0.007% gaps.

The ArrowGrid parallel wrapper (<https://github.com/skoren/ArrowGrid>) was used for running the Arrow consensus framework (<http://github.com/PacificBiosciences/GenomicConsensus/>) (Chin et al., 2013) within the PacBio SMRT Analysis Software to polish the gap-filled assemblies using the PacBio reads. Details on the original consensus model used for polishing can be found in Chin et al. 2013. Following Arrow polishing, the PilonGrid parallel wrapper (<https://github.com/skoren/PilonGrid>) (Walker et al., 2014) was used for running Pilon polishing (Walker et al., 2014) using the Illumina 2x250 genomic reads.

Following the gap filling and polishing steps, we sought to determine whether all transcripts in our independently generated transcriptomes were represented in our primary assemblies or whether some sequences had been filtered into the secondary assemblies. All transcripts from the transcriptome were aligned to the primary and secondary sets using the alien index and any transcript that had a better alignment to the secondary set was added back to the primary set, making the final size of the primary set for *H. symbiolongicarpus* 406.693 Mbp. We only added partial scaffolds back to the primary set to avoid increasing the amount of duplicated sequence. We kept the complete scaffolds in the secondary set, including the partial scaffold

sequence that we added to the secondary set, so there is some redundancy between the primary and secondary sets for these sequences.

### **5.6.2 RNA extraction and sequencing**

RNA extraction and sequencing was performed as previously described (2.6.2).

### **5.6.3 Assembly of the 291-10 ARC and alignment to the ARC-F reference**

The 291-10 ARC was assembled and aligned to the ARC-F reference sequence, obtained in 2.0, as described in 2.6.3. The contigs from the 291-10 assembly that aligned with the ARC-F (representing both haplotypes) were aligned to one another to verify the genomic architecture.

### **5.6.4 Annotation of *Alr* genes in the 291-10 ARC**

*Alr* genes were annotated in the 291-10 ARC as previously described (2.6.4). One update was made to the pipeline that involved the database used for the BLAST results. The database used to query the DNA segments was updated to include all the new *Alr* gene annotations from the ARC-F so that the homologous genes could be more easily identified and any *Alr*-like genes with more variation from the original dataset could be identified. All BLAST results for *Alr*-like genes were found on the contigs that had aligned to the ARC (5.6.3).

### 5.6.5 Variation analysis of alleles

The peptide sequences were downloaded from the *Alr*-like annotations made in WebApollo. For gene models with alternative splice variants, only equivalent splice variants were compared when counting polymorphic sites. The shorter splice variants did not provide any additional insights into the level of polymorphism between alleles. Thus, only the first, or full-length, splice variant was included in the analysis results. For pseudogene models that were putative or bona fide in other alleles, only the longest in-frame sequence was used to determine variation between alleles, if available. Jalview was used to manipulate and compare *Alr* sequences.

## 6.0 Discussion and Future Directions

### 6.1 The ARC contains a large *Alr*-like gene family that are novel members of the IgSF

Prior to this work, only *Alr1*, *Alr2*, and some candidate *Alr*-like sequences directly surrounding them were known to be present in *Hydractinia*. *Alr1* and *Alr2* both impact the allorecognition response (Cadavid et al., 2004) and were shown to have differential effects on the allorecognition phenotype (Powell et al., 2007) in defined genetic lines. However, it was observed that a colony's genotype at *Alr1* and *Alr2* can fail to predict allorecognition in some cases, particularly between outbred colonies (Cadavid et al., 2004; Powell et al., 2007; Nicotra et al., 2009; Rosa et al., 2010). These observations suggested that there may be more allodeterminants in *Hydractinia* that may have similarities to *Alr1* and *Alr2*. However, testing this hypothesis would have been difficult using the same mapping strategy likely because a haplotype which shows only the effects of a third allodeterminant may not be achievable by simply crossing different strains especially considering the differential effects on phenotype observed in *Alr1* and *Alr2*. The sequencing of the *Hydractinia* genome enabled the comprehensive study of the ARC which I showed in Chapters 2.0, 3.0, and 5.0.

To determine how many additional *Alr*-like sequences are present in the *Hydractinia* genome, I annotated and compared the *Alr* content from a colony which contained the haplotype used in mapping studies (haplotype F) and two outbred haplotypes (haplotypes pr, se). This resulted in the identification of 41 *Alr* loci in the F haplotype and a total of 56 unique loci in the pr and se haplotypes. All *Alr* sequences shared a similar domain architecture with most *Alr* sequences containing a similar domain set as *Alr1* (Figure 22). Despite the similar domain

architecture, most of these genes were vastly different at the sequence level with an average percent identity between any two Alr protein sequences (excluding splice variants of the same gene) of  $24.3\% \pm 8.6\%$  (Figure 19). This suggested that the *Alr* gene family may have been the result of multiple duplications and a substantial amount of sequence evolution to generate its diversity.

Following the identification of so many additional *Alr* loci, the increased sequence variation among them provided a diverse dataset with which to search for homology. I searched for deep sequence homology using HMMER and HHpred and predicted each domain's structure using the *de novo* structure prediction algorithm AlphaFold. These methods together confidently places the Alr domains as novel members of the V-set (domain 1), I-set (domain 2 and domain 3), or fibronectin III-like (ECS) Ig families despite lacking many canonical motifs. Nearly all representative domains in these PFAM datasets come from vertebrate models which cannot account for the evolutionary history of these domains in other model systems. The canonical Ig domain motifs in the PFAM dataset were replaced by unique sequence signatures in the Alr domains that maintain the same structural fold based on my analysis (3.4.1, 3.4.2, 3.4.3). This is quite remarkable and reinforces my confidence that these are the correct domain folds found in Alr proteins. These unique sequence signatures found in the Alr domains showcases an opportunity to study distant evolutionary relationships of Ig domains.

The phylogenetic distribution of the *Alr* gene family and *Alr*-based allorecognition among cnidarians is unknown. Nonetheless, the organization of the ARC reinforces similarities between *Hydractinia* and other species in which allorecognition is controlled by genomic clusters of related genes (Grice et al., 2017). 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 all of these systems may become possible in the future as others and I survey a range of metazoan genomes and simultaneously deepen the molecular understanding of how invertebrate allorecognition works.

## **6.2 Alr2 binding specificity can evolve quickly with point mutations in domain 1**

In the study showing *Alr1* and *Alr2* could bind homophilically (Karadge et al., 2015), nearly all isoforms contained numerous differences between their ectodomains (>20) and were capable of binding in an isoform-specific manner. In only one case, a pair of *Alr1* isoforms that had only four differences occurring in domain 2 and the FN3-like domain did not bind isoform-specifically. This suggested that not all amino acid changes are capable of altering the binding specificity either because of their location in the domain or because there are not enough mutations to sufficiently alter binding specificity. Determining which residues control the binding specificity when testing isoforms with numerous differences is challenging due to the number of pairwise combinations that would need to be tested in order to eliminate residues that do not affect the binding specificity. Thus, finding such a closely related set of isoforms was crucial to beginning to understand how binding specificity can evolve. Most systems that study binding specificity compare sequences with several differences between them or that are chimeric and stochastically test residue changes. *Alr2* is one of the most polymorphic genes reported in any organism and offers a huge advantage as there are numerous pairs of sequences with very few differences between them that are present and function in wild populations. These alleles detail the sequence evolution that has occurred with a high degree of certainty. This drastically improves the

confidence in results related to binding specificity that can be correlated with the evolutionary history of the alleles since they can be traced.

In Chapter 4.0, I identified the three residues (N32Y, S44G, G47E) which caused shifts in binding specificity in six closely related isoforms. I showed that specificity can change in as few as one change (N32Y) or cumulatively (as with S44G and G47E). While every change will not have an equivalent effect on the binding specificity, these results put into perspective how many potentially unique binding specificities may exist in the hundreds of isoforms of *Alr2* surveyed in nature thus far (Gloria-Soria et al., 2012) and is consistent with the rates of histocompatibility found in nature (Grosberg, 1988; Nicotra & Buss, 2005).

All the isoforms that I tested, as well as all isoforms tested to date, were capable of homophilic binding. It is yet to be determined whether there are alleles of *Alr2* (or any other *Alr*) that exist in nature which do not exhibit homophilic binding. Since *Alr2* is involved in allorecognition, it is possible that *Alr2* alleles are under strict selection which eliminates the presence of many alleles encoding non-functional (non-binding) proteins. In some preliminary work, I tested whether I could alter specific residues to disrupt the *Alr2* binding function. In several cases, the expressed *Alr2* protein was not trafficked to the surface and may have not been properly processed leading its accumulation intracellularly (data not shown). This was likely the result of the mutation selected, but also highlighted that more information regarding the molecular mechanism of interaction should be obtained before stochastically mutating residues.

One follow-up to this work that was not explored was the effects on binding specificity caused by the I-set domains (domain 2 and domain 3) and the fibronectin III-like domain (ECS). It is currently unknown which domains interact in *trans*. Changes in domain 1 are sufficient to disrupt binding between otherwise compatible isoforms, as shown in chapter 4.0, which suggests

that domain 1 directly binds in *trans* in some orientation (Figure 57). However, the interactions of the membrane proximal domains are unknown. Similar to the approach in chapter 4.0, alleles with changes localized in only the I-set or fibronectin III-like domains could help elucidate their function in homophilic binding.

### **6.3 *Alr1*, *Alr2*, and *Alr6* are highly polymorphic**

All of the *Alr* genes were less polymorphic than *Alr1* and *Alr2* based on the three haplotypes compared. However, this does not exclude the possibility that *Alrs3-38* could have as high or even higher rates of polymorphism in other haplotypes or in other populations of *Hydractinia*. The less polymorphic *Alrs* may have functions aside from determining recognition specificity. Given their conserved domain architecture, it is possible other *Alr* genes are expressed and translated into protein and play a role in cell-cell adhesion. The *Alr* genes may also be a source of sequence diversity for those genes which determine allorecognition specificity. For example, *Alr1-pr* and *Alr1-se* were shown to encode very dissimilar domain sequences that were very similar to *Alr18-pr* and *Alr19B-se*, respectively, suggesting that exon shuffling occurs between *Alr* genes (Figure 64). The presence of so many *Alr*-like genes, especially within compact regions such as the clusters, suggests that gene duplication may be a frequent event in the ARC and would be a prime system for neofunctionalization. Given the appropriate circumstances, the function of *Alr1* and *Alr2* in allorecognition, beyond the purpose of a cell adhesion molecule, could have been the result of neofunctionalization.

The only gene aside from *Alr1* and *Alr2* that exhibited high levels of polymorphism was *Alr6*. In many self/non-self recognition systems, as is the case in the *Hydractinia* allorecognition

system, polymorphism is often a key trait that determines recognition specificity. *Alr6* was present in the region that was homogenized in the lab generated lines (Figure 70). It is highly probable that this homogenized region contains the third allodeterminant as its effect on allorecognition was masked when comparing the phenotypes between the R and F haplotypes. If the third allodeterminant is also highly polymorphic like *Alr1* and *Alr2*, *Alr6* would perhaps be the strongest candidate gene identified as it resides in Cluster A and is the next most polymorphic. *Alr6* is also interesting in that it encodes three splice variants. Two of these resemble splice variants found in *Alr1* (Figure 17). These similarities may be purely coincidental. However, if these similarities have any significance, several observations can be made. First, alternative splice variants may be important for allorecognition genes involved in specificity. Both *Alr1* and *Alr2* encode splice variants. While the purpose of those splice variants is unknown, it is possible that there may be some function that has not been observed yet. Second, *Alr1* and *Alr6* are both localized at one end of the clusters they are in and have some of the same splice patterns. This could suggest that *Alr1* and *Alr6* are related to one another, possibly through a gene duplication event, after which exon shuffling and high rates of mutation led to the unique coding sequences which no longer share significant sequence homology. While none of these observations regarding *Alr1* and *Alr6* can be tested currently, it highlights the complexity of the allorecognition system and events which led to the presence of so many *Alr* sequences which should be considered in all cases as the evolutionary history of the family is explored.

## 6.4 Future directions

To date, RNA can only be reliably extracted from polyp tissue. Extracting good quality RNA from the mat and stolon tissue – where allorecognition responses occur – has not been achieved yet. While the reason for low quality RNA remains unknown, a possible factor that makes extracting high quality RNA difficult may be the chitinous layer that grows between the epithelium of the mat and the surface it grows on. Attempts have been made to collect only the mat tissue and leave behind any chitin. However, this has not solved the low RNA quality issue. When a method is developed to extract high-quality RNA reliably from the mat tissue, RNA-seq could be used to estimate the expression level of *Alrs* more accurately in these tissues. The *Alr* expression levels, as well as localization, in the mat tissue are currently unknown. In addition, collecting RNA at various timepoints during an allorecognition response will help us understand their expression during fusion, transitory fusion, or rejection and may guide us to other allorecognition-related genes.

While homophilic binding has only been tested with Alr1 and Alr2 *in vitro*, the conserved domain architecture among the Alr family suggests that the new Alr proteins may also be capable of extracellular binding or cell adhesion-like interactions. Based on Alr1 and Alr2, other Alr proteins may also function through the same homophilic binding mechanism, however it is also possible that they may interact through alternative mechanics such as heterophilic binding (e.g. Alr3 with Alr4). In future work, cloning multiple alleles of these new *Alr* genes and testing them *in vitro* would help answer these questions.

The combined sequence and structural analysis of the Alr domains produced results that place the Alr family as novel members of Ig domains. While I am confident in the structural predictions produced by AlphaFold, obtaining additional data from crystallized structures would

increase my confidence in the proper orientation of the backbone and side chains of the domain. Crystal structures may improve the chances of modeling the homophilic binding interaction and aid in identifying potential residues that may change the binding specificity. This would help focus the selection of *Alr* sequences to test experimentally and determine the effect of individual residues on binding specificity. Several attempts have been made to crystallize Alr1 and Alr2 to no avail. From attempting to crystallize the entire ectodomain to crystallizing even a single domain, Alr1 and Alr2 have proven resistant to the most commonly used crystallization conditions. This may be due to the proteins being unstable when expressed in solution and not membrane bound. Current efforts are focused on crystallizing the Alr2 isoforms from Chapter 4.0. When obtained, the crystal structures could be directly analyzed in combination with the binding specificity data which would quickly improve our understanding of the homophilic binding mechanism.

While the identity of *Alr1* and *Alr2* can predict the alloresponse outcome in select cases, the signaling mechanism which drives histocompatibility responses between colonies is not clear. The progression of any alloresponse in *Hydractinia* starts with the migration of nematocytes to the border of contact. After some time, if colonies are compatible and can fuse, nematocytes will migrate away from the border. If, however, the colonies are incompatible, the nematocytes will discharge which is the ultimate outcome of a rejection response. The alloresponse only occurs between *Hydractinia* and does not occur when it grows into contact with other species. Given the initial migration of nematocytes regardless of the compatibility outcome, it has been hypothesized that *Hydractinia* may contain a marker which initiates the alloresponse. Following a compatible reaction between at least Alr1 and Alr2, the alloresponse would shift to a fusion outcome. Some Alr proteins appear to have an ITIM or ITAM in their cytoplasmic tail that may be involved in signaling their compatibility after contact has been made. For example, Alr2 contains an ITIM in

its cytoplasmic tail. A proper Alr2 interaction could result in the ITIM interacting with intracellular proteins which eventually inhibit the signaling related to activating the rejection response. Identifying potential binding partners to the cytoplasmic tail could help identify any conserved signaling pathways which have been maintained in immune-like molecules.

The correlation between binding specificity and allorecognition specificity has been tested in some *Alr1* and *Alr2* alleles (Cadavid et al., 2004; Nicotra et al., 2009; Rosa et al., 2010; Karadge et al., 2015) however it has not been exhaustively tested with all known *Alr1* and *Alr2* alleles. While it is likely that isoforms which bind lead to fusion, what about isoforms which weakly interact with each other? For example, some of the isoforms which I tested with Alr2-214E06 in Chapter 4.0 appeared to exhibit weak heterophilic interactions. Would these weak interactions be enough to initiate or contribute to a histocompatible response or would they be interpreted as incompatible? The only way to determine this *in vivo* would be to test colonies encoding the allele pairs which exhibit this *in vitro* binding phenotype. Part of the difficulty in examining this principle *in vivo* is that the genotypes of each allodeterminant required to compare binding specificity with allorecognition specificity do not currently exist in sampled colonies and would require an intensive breeding program to develop if even possible. In addition, the complexity of the system involving at least three allodeterminants could also make it difficult to isolate phenotypic differences with binding specificity differences of each of the allodeterminants. The recent expansion of genomic tools available to *Hydractinia* to include gene editing techniques (Sanders et al., 2018) provides an alternate approach for addressing how Alr binding specificity and allorecognition specificity are related in *Hydractinia*. While the process to successfully edit *Hydractinia* may be more direct than the breeding required to generate colonies to test the right

allele pairs, such an approach would require months to generate and verify that the encoded alleles are properly inserted and exclusively present (i.e. present in both copies).

The cell aggregation assay was used to determine whether or not the Alr2 isoforms I tested were capable of binding. The observation of the weak heterophilic phenotype between 214E06 and some of the other isoforms indicated that a more quantitative approach would be necessary in order to fully understand Alr interactions. During the development of the assay, positive controls were used from well-known homophilic binding proteins, such as N-cadherin (Katsamba et al., 2009), and resulted in extremely large aggregates with thousands of cells as opposed to the hundreds of cells observed in Alr-presenting aggregates. Given the qualitative nature of the assay, it cannot be determined how different the binding interaction strength is between various isoforms. However, the consistent observation of Alr-presenting cells forming aggregates smaller than those presenting a homophilic binding protein which has been characterized to be a strong interaction suggests that Alr2 is a relatively weaker homophilic binding protein which makes interpreting the results of a qualitative assay more difficult given the small size of the aggregates. This again highlights the need to develop of a quantifiable assay for testing these interactions as well as quantitatively measuring the interaction to be able to compare it with the strength of other homophilic binding interactions. One such approach which I attempted to develop in the lab was an ELISA-based binding assay (Wojtowicz et al., 2007). This assay was developed to be used as a high-throughput approach testing numerous Dscam isoforms from *Drosophila* that encoded Ig domains. The assay utilized alternately tagged ectodomains of Dscam in order to capture and detect interactions between isoform pairs. The assay required two antibodies, one of which is no longer commercially produced. Without a reliable source for one of the antibodies, developing this assay further was postponed until a reliable source could be found or an alternative tag could be used.



However, this ELISA-based approach would significantly improve the number of alleles we could test in pairwise as well as allow us to quantify the interaction strength between pairs.

The presence of the six perfectly conserved cysteines in the I-set and FnIII-like domains is a unique characteristic of the Alr domains. Based on their predicted structure, nearly all tandem domains had the interdomain disulfide bond in addition to intradomain disulfide bonds present in each domain (Figure 39). The intradomain disulfide bonds occurred between strands on the same face of the domain rather than between opposing stands which would be expected for increased domain stability. The interdomain disulfide bond if formed may be important for keeping the orientation between the I-set and FnIII-like domain. Whether any of these disulfide bonds are formed *in vitro* (and *in vivo*) is not yet determined. If some of these cysteines are not involved in an intramolecular disulfide bond, they may be free to form an intermolecular bond thus stabilizing protein dimerization on the cell surface. An approach to address whether disulfide bonds are formed between each of these would include mass spectrometry. Given the sequence variation present in the Alrs, each of the disulfide bonds could be identified separately within the same sample. Alternatively, a pull-down assay with the right extraction method may also be a viable approach to detecting whether Alrs form dimers on the cell surface. Interestingly, the possibility for Alrs to dimerize would increase the avidity of their interactions between cell surfaces making up for the relatively “weak” interactions observed from the cell aggregation assay.

It is unclear whether the new *Alrs* identified in Chapters 2.0 and 5.0 play a biological role in allorecognition or in some other context for *Hydractinia*. While it is anticipated that at least one additional *Alr* may be involved in determining specificity, the presence of so many other *Alrs* which do not appear to impact the allorecognition response suggests they have another role. In addition, nearly all other *Alrs*, aside from *Alr6*, are significantly less variable than *Alr1* and *Alr2*

in the haplotypes available. While this does not exclude the possibility that they are also highly polymorphic when analyzed in other haplotypes, their low sequence polymorphism may be an important feature for how they function, which may not be determining histocompatibility. The localization of all Alr proteins, including Alr1 and Alr2, has not yet been explicitly shown. While the expected localization of Alr1 and Alr2 is the cell surface since that is how it behaves *in vitro*, where it is expressed within the animal has not been definitively shown. Antibodies could be used to localize where Alr1 and Alr2 (as well as any other Alrs) are present in *Hydractinia*. However, to date no Alr-specific antibodies have been able to be produced. In Karadge *et al* (2015), Alr1 and Alr2 were shown to be expressed on the cell surface through the use of a FLAG tag and anti-FLAG antibody. With the recent achievement of gene editing in *Hydractinia* (Sanders et al., 2018), it has become possible to tag genes of interest with fluorescent reporters. It is also possible that genes could be tagged in a manner similar to the *in vitro* constructs used which would allow them to be localized with antibody staining in *Hydractinia*.

## **7.0 Externship research: Cryopreservation of *Hydractinia symbiolongicarpus* sperm to support community-based repository development for the preservation of genetic resources**

### **7.1 Foreword**

The work in this chapter was collected during my ISB program externship with the Aquatic Germplasm and Genetic Resources Center (AGGRC) at Louisiana State University. This work will be adapted into a manuscript for publication in which I am first author: Aidan L. Huene, Matthew L. Nicotra, Virginia M. Weis, Terrence R. Tiersch (Unpublished, 2022).

### **7.2 Summary**

*Hydractinia symbiolongicarpus* is an emerging model organism in which cutting-edge genomic tools and resources are being developed for use in a growing number of research fields. One limitation of this model system is the lack of long-term storage for genetic resources. The goal in this study was to establish a generalizable approach to sperm cryopreservation that would support future repository development for *Hydractinia* and any other communities seeking to establish long-term storage options. Our approach was to: 1) Assess sperm characteristics and standardize collection and processing; 2) Assess acute toxicity to cryoprotectants, and 3) Evaluate and refine freezing conditions to permit post-thaw fertilization and produce viable offspring. By following this approach, we quickly developed a protocol which incubated *Hydractinia* sperm in

5% DMSO, equilibrated at 4°C for 20 min, and cooled at a rate of 20°C/min to -80°C at a cell concentration of  $10^8$ - $10^9$ /mL in 0.25-mL aliquots. These aliquots were able to fertilize 150-300 eggs that yielded offspring that could metamorphose into juvenile polyps. The success achieved with the quickly developed protocol leaves much room for improvement and comparison between various cryoprotectants in future work.

### 7.3 Introduction

*Hydractinia symbiolongicarpus* is a colonial cnidarian and an established model for evolutionary developmental biology, stem cell biology, regeneration, and allorecognition (Günter Plickert et al., 2012; Gahan et al., 2016; Nicotra, 2019). In recent years, efforts to improve *Hydractinia* as a model system have included generation of robust laboratory strains for use by the research community, sequencing of these strains through the *Hydractinia* Genome Project (Schnitzler et al., 2017), and establishment of methods to produce transgenic animals via the random integration of exogenous DNA (Bradshaw et al., 2015) or targeted integration via CRISPR/Cas9-mediated gene knock-in (Sanders et al., 2018).

An increasing limitation to the expanded use of *Hydractinia* as a model is the lack of long-term storage options for genetic resources. Over the years, laboratories have collected and bred hundreds of genotypically distinct colonies, while simultaneously generating strains bearing various transgenes. In all cases, these animals have had to be maintained as live animals vulnerable to accidents, disease, and improper handling, which can result in death and permanent loss of genotypes. While *Hydractinia* colonies can be maintained for decades under laboratory conditions, it is increasingly costly in terms of labor and space.

To address this limitation, we sought to evaluate the feasibility and potential utility of cryopreservation as an archival storage method. As an immediate benefit, cryopreservation would allow “backing-up” animals that are valuable genetic resources. In addition, as a long-term benefit beyond laboratory use, cryopreserved stocks would allow user groups from across the research community to store and access samples on demand rather than requiring time and resources to grow or collect new animals. While the ultimate goal would be cryopreservation of germplasm and somatic tissues from all life stages, here we focused on *Hydractinia* sperm as the most amenable to cryopreservation based on previous success in corals (M Hagedorn et al., 2006; Mary Hagedorn, Oppen, et al., 2012; Mary Hagedorn et al., 2013, 2019) and the anemone *Nematostella* (Matt Gibson and Shane Merryman, personal communication).

Although much is known about *Hydractinia* embryonic development and the differentiation of *Hydractinia* germ cells (Weis et al., 1985; Mali et al., 2011; Kraus et al., 2014), much less is known about *Hydractinia* germplasm after its release, beyond what is necessary for routine breeding. It is well established that *Hydractinia* are dioecious and have gonozooids (reproductive polyps) that bear multiple gonophores (gamete-filled structures) that release either sperm or eggs. Healthy *Hydractinia* release gametes daily. Researchers typically allow male and female colonies to spawn together in the same water or they collect eggs and sperm separately and then mix them within 30 min of spawning. Anecdotal evidence suggests waiting longer than 30 min decreases the quantity and quality of embryos.

After fertilization, each embryo develops into planula larva (1-4 d) before permanently attaching to the surface and metamorphosing into a juvenile primary polyp. The animal then grows by extending structures called stolons across the surface, from which additional polyps are produced to create a colony. Colonies become sexually mature within 1-2 months. Under

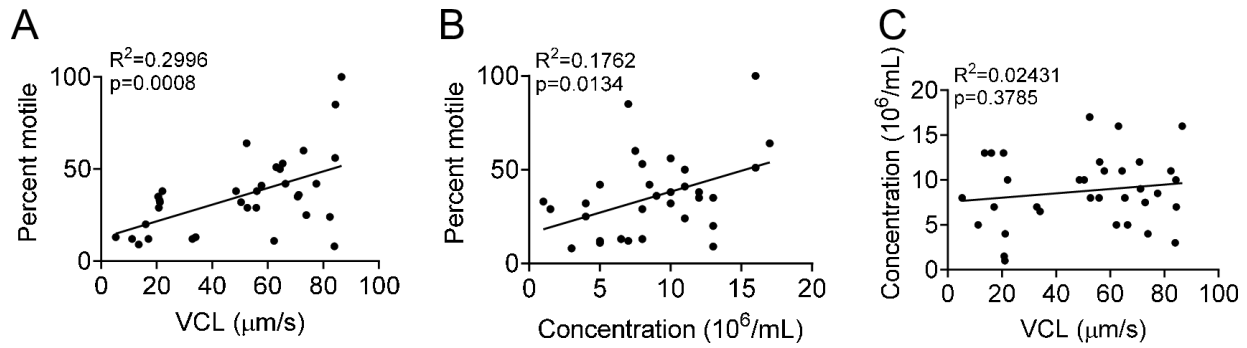
laboratory conditions the number of offspring that are male or female has been observed to be consistent with a 1:1 sex ratio.

Successful cryopreservation of sperm cells requires the balance of multiple parameters (Tiersch, 2011a). These include the storage temperature, the temperature difference and time that elapses between sperm collection and freezing, sperm concentration at the time of freezing, choice and concentration of cryoprotectant, cooling method and rate, thawing method and rate, and the conditions under which thawed sperm will be used for fertilization (Torres & Tiersch, 2018). Here we detail a systematic three-part approach to: 1) determine basic characteristics of *Hydractinia* sperm and standardize collection and processing; 2) test the toxicity of commonly used cryoprotectants, and 3) identify conditions that maximize the likelihood of cryopreserved sperm samples being capable of fertilization after thawing.

## **7.4 Results**

### **7.4.1 Sperm motility and viability**

To assess sperm characteristics, we measured sperm from 35 clouds (each collected in 10  $\mu$ l) using CASA. Mean velocity was  $50.8 \pm 26.2$   $\mu$ m/s, mean percent that were motile was  $37 \pm 22\%$  and mean concentration was  $9.37 \pm 5.31 \times 10^6$ /mL. Based on linear regression, velocity and motility were correlated ( $R^2 = 0.2804$ ,  $P = 0.0011$ ) (Figure 71A), as were concentration and motility ( $R^2 = 0.2870$ ,  $P = 0.0009$ ) (Figure 71B). Velocity and concentration were not correlated ( $R^2 = 0.02365$ ,  $P = 0.3778$ ) (Figure 71C).



**Figure 71. Correlations among sperm velocity, motility, and concentration.**

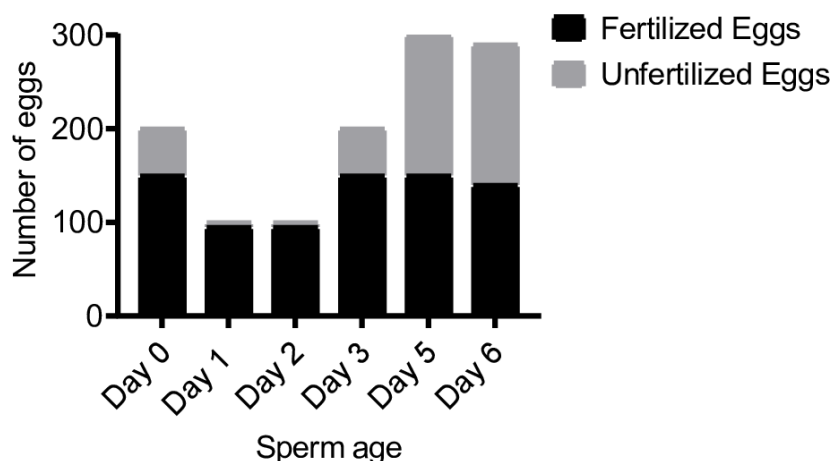
Each point represents a sperm cloud (N = 35). (A) Distribution of sperm based on velocity and the number motile.

(B) Distribution of sperm based on concentration and number motile. (C) Distribution of sperm comparing velocity and concentration.

To determine the effect of temperature on sperm viability, we compared the motility of freshly collected sperm held at room temperature (22°C) to that of sperm held in a 4°C refrigerator. At room temperature, the number of motile sperm declined over 6 hr, such that by 7 hr only twitching was observed (tail movement without progressive motility). In contrast, sperm kept at 4°C retained progressive motility 7 hr after collection, although the total number of motile sperm and the velocity visibly decreased. By 23 hr, no sperm were motile, but approximately 40% assessed manually were still twitching. Thus, holding sperm at 4°C prolonged motility.

The observation that sperm held at 4°C were still moving after 23 hr raised the question of whether they could still fertilize eggs and, if so, whether sperm would remain viable after longer storage times. To address this question, we collected ~150 sperm clouds and used the sperm to fertilize freshly collected eggs over the following 6 d (Figure 72). We performed daily routine breeding to serve as a positive control for egg fertilization; nearly all of the eggs (>95%) were fertilized each day indicating that there was no appreciable differences in egg quality for fertilization. On day 0, we mixed  $2 \times 10^7$  sperm (3 mL) with ~200 eggs, which resulted in ~150

embryos. Because ~50 eggs remained unfertilized, we interpreted this to indicate that the defined sperm number in this sample ( $2 \times 10^7$ ) were capable of fertilizing ~150 eggs.



**Figure 72. Estimated sperm fertilization capacity over time.**

Each day,  $2 \times 10^7$  sperm cells from the same collection aliquot were used to fertilize the freshly collected eggs in 30 mL FSW. On Days 1 and 2, only ~100 eggs were available for exposure to sperm. On the other days, a surplus of eggs were collected for exposure.

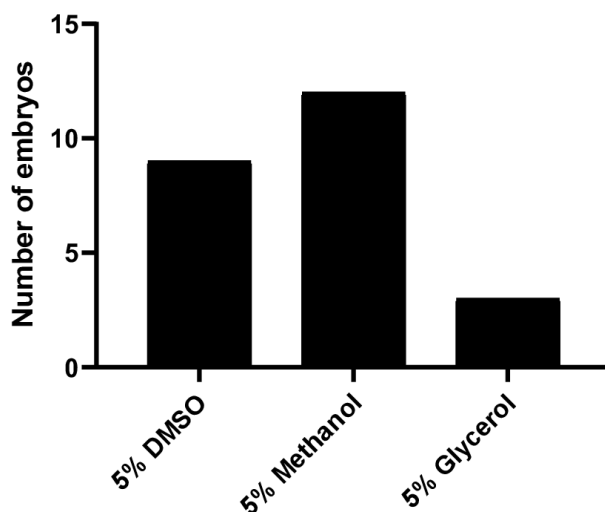
On each subsequent day, we mixed the same amount of stored sperm with as many eggs as we could collect and estimated the total number of fertilized and unfertilized embryos. We found that  $2 \times 10^7$  sperm consistently fertilized ~150 eggs after 3, 5, and 6 d at 4°C. On days 1 and 2, we were only able to collect ~95 eggs, nearly all of which were fertilized. These latter data were consistent with the notion that  $2 \times 10^7$  sperm could fertilize ~150 eggs. In these experiments, all embryos developed and metamorphosed into normal juvenile colonies.



### 7.4.2 Determining cryoprotectant toxicity to sperm

We tested the acute toxicity of three common cryoprotectants (DMSO, methanol, and glycerol). Sperm incubated with the three concentrations (5, 10 and 15%) of DMSO or methanol displayed comparable motility after 30 min. In contrast, sperm exposed to 15% glycerol ceased moving immediately, while those exposed to 10% and 5% glycerol were non-motile within 30 min.

To determine whether cryoprotectant-treated sperm would be able to fertilize eggs, we exposed sperm to each cryoprotectant for 30 min and mixed  $4.1 \times 10^6$  sperm with 40 freshly collected eggs in a total volume of 50 mL. Sperm exposed to 10% or 15% of any cryoprotectant were unable to fertilize eggs. In contrast, sperm treated with 5% of any cryoprotectant yielded 3-12 embryos (Figure 73). From this, we concluded that 5% DMSO or methanol would be suitable cryoprotectants.



**Figure 73. Number of fertilized eggs using cryoprotectant-treated sperm.**

For each condition, 30-40 eggs were exposed to  $4.1 \times 10^6$  sperm in a total volume of 50 mL.

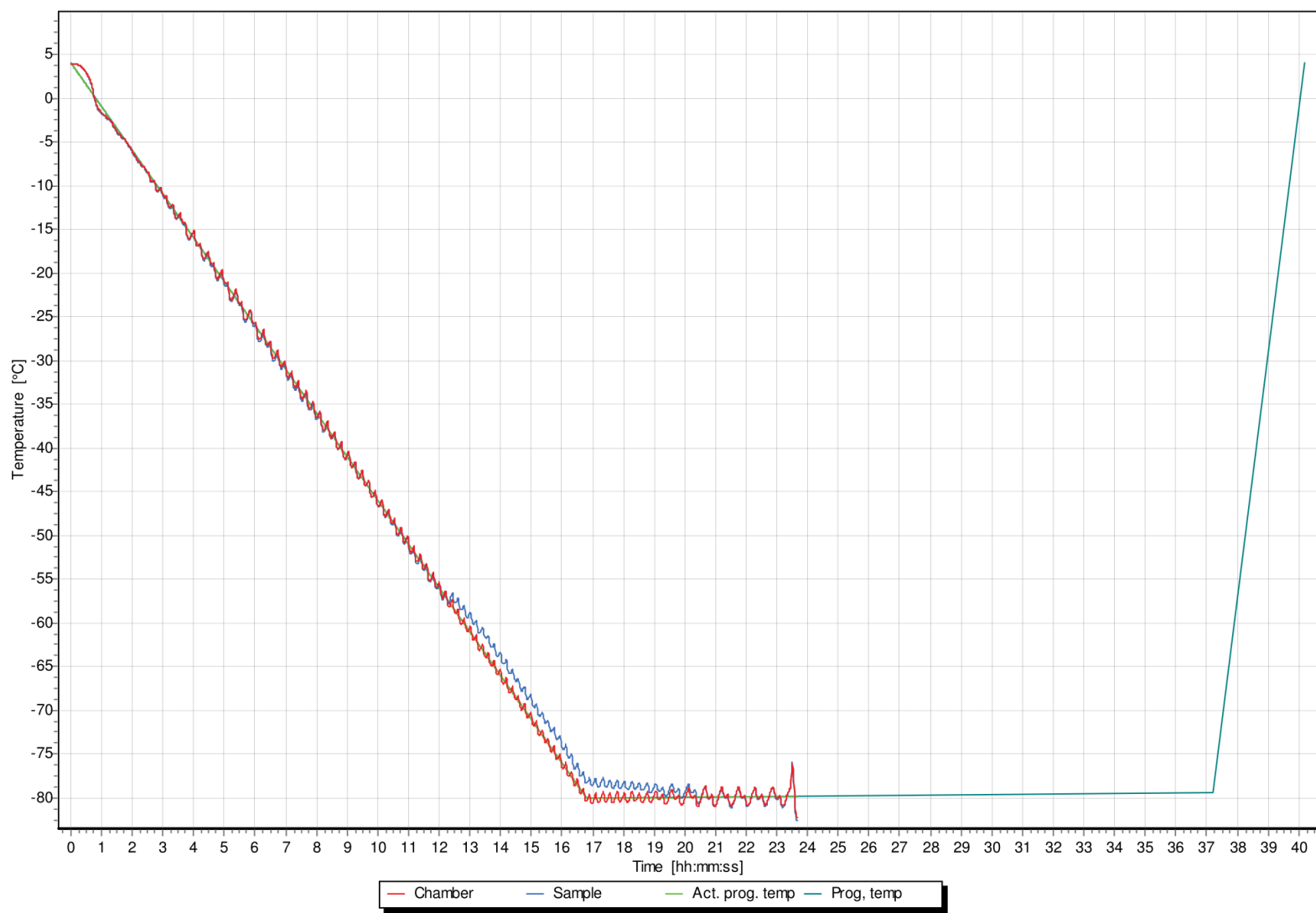
### 7.4.3 Identifying suitable freezing conditions

While many factors affect the quality of cryopreserved sperm, three key parameters must be balanced: cryoprotectant concentration, sample concentration, and cooling rate. For example, higher cryoprotectant concentrations can be more toxic, whereas lower concentrations may not sufficiently protect the cells. Moreover, the toxicity of a given concentration of cryoprotectant often decreases as the sample concentration increases (Tiersch, 2011a). The cooling rate must also be slow enough to allow cells to dehydrate sufficiently (to minimize intracellular ice formation), but fast enough to freeze them before concentrations of intracellular salts or pH (i.e., solution effects) or the cryoprotectant become damaging.

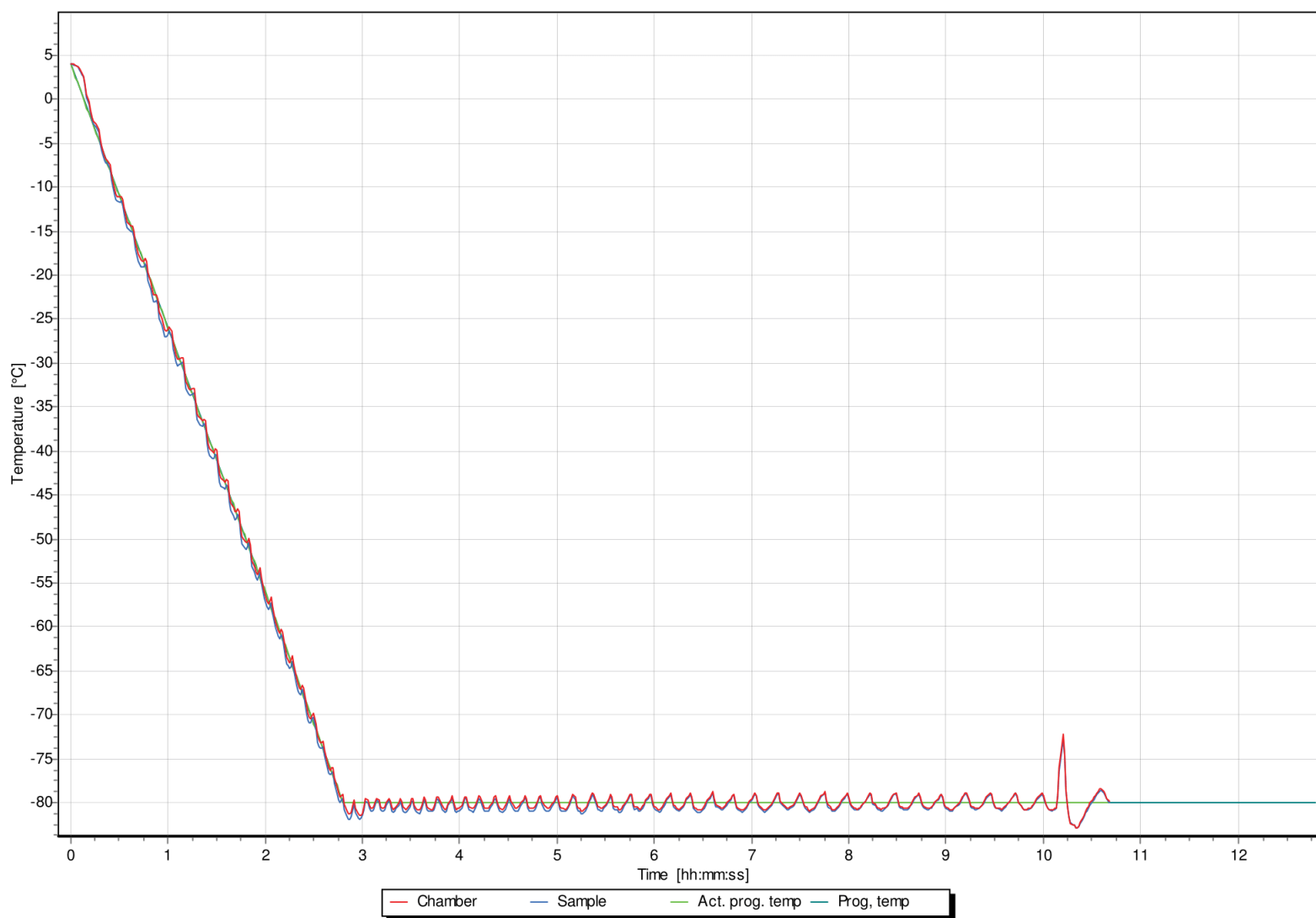
To survey the effects of freezing rate on sperm in either 5% DMSO or 5% methanol, we cooled sperm at 5°C/min and 30°C/min (Table 20, Experiment 1; Figure 74 and Figure 75). Samples were stored in liquid nitrogen for at least 21 hr before they were thawed and evaluated. In all conditions, the concentration of intact sperm in the thawed samples was reduced from  $1 \times 10^7$  to  $2 \times 10^6$  or fewer, nearly ten-fold, likely due to cell rupture either during freezing or thawing. Overall, between  $5 \times 10^4$  and  $1 \times 10^5$  fewer sperm were detected in the 30°C/min samples than in the 5°C/min samples suggesting that the faster rate did not allow sufficient osmotic egress and intracellular ice was formed. We incubated aliquots of each thawed sample with 75 freshly collected eggs. Despite the low numbers of sperm used ( $\leq 5 \times 10^5$ ), at least one egg was fertilized in each condition. This indicated the presence of viable sperm and suggested that increasing the effective sperm concentration would increase fertilization.

**Table 20. Overview of frozen samples and fertilization potential.**

	Experiment 1				Experiment 2
Cryoprotectant	5% DMSO	5% DMSO	5% Methanol	5% Methanol	5% DMSO
Initial sperm concentration (sperm/mL)	$1 \times 10^7$	$1 \times 10^7$	$1 \times 10^7$	$1 \times 10^7$	$5 \times 10^7$
Cooling rate (°C/min)	5	30	5	30	20
Hours stored frozen	69	69	69	69	21
Thawed sperm concentration (sperm/mL)	$2 \times 10^6$	$1.5 \times 10^6$	$2 \times 10^6$	$1 \times 10^6$	$5 \times 10^7$
Total sperm mixed with 75 eggs	$5 \times 10^5$	$3.8 \times 10^5$	$5 \times 10^5$	$2.5 \times 10^5$	$1.2 \times 10^7$
Number of embryos	2	2	2	1	10

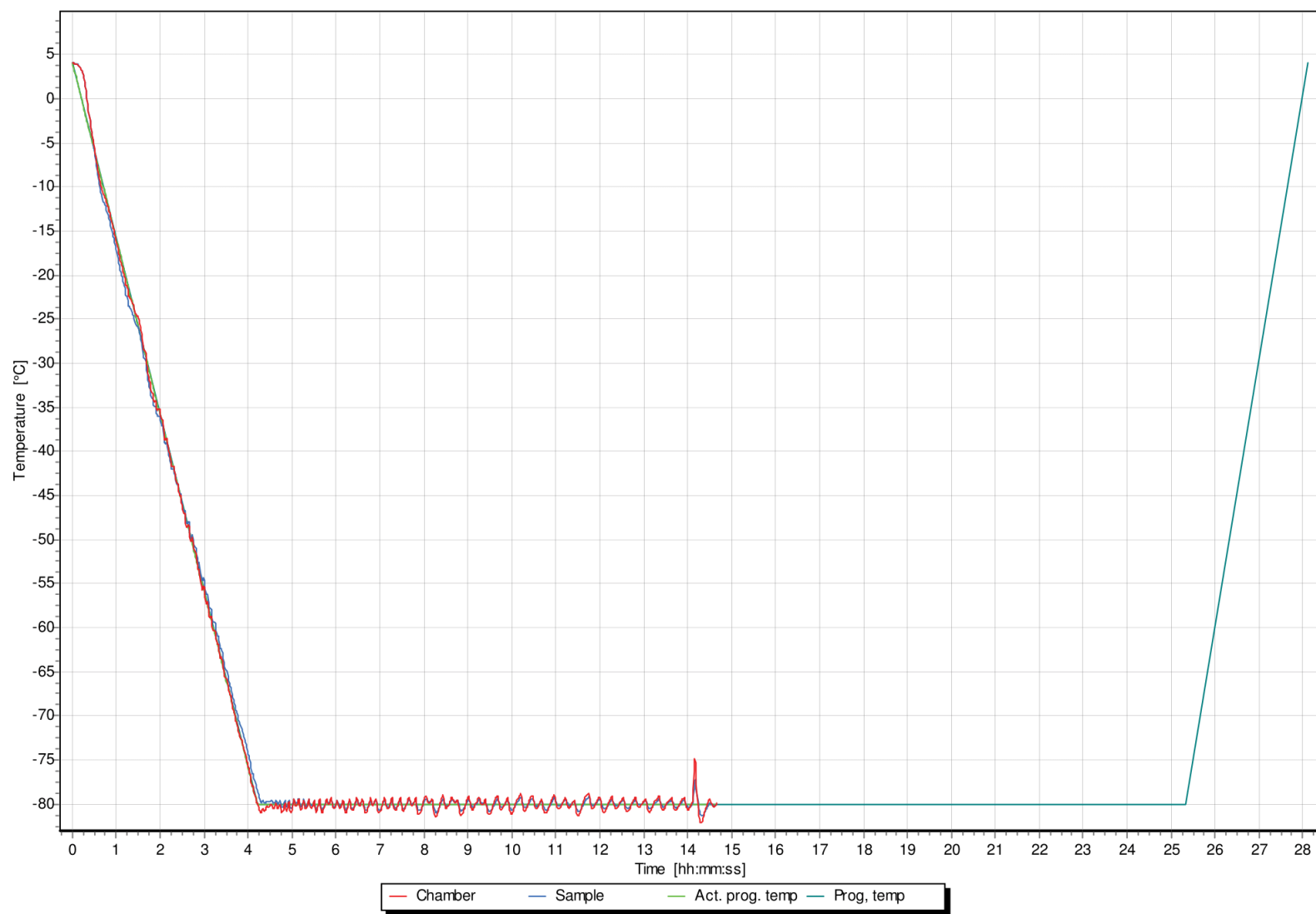


**Figure 74. Cooling curve for Experiment 1, 5°C/min.**



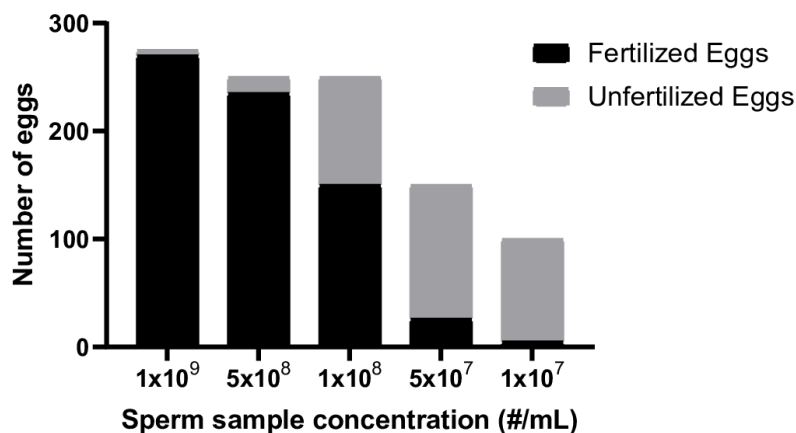
**Figure 75. Cooling curve for Experiment 1, 30°C/min.**

We increased the volume and concentration of sperm collected by fabricating a sperm collection chamber by 3-D printing that allowed incubation of as many as ten slides bearing male colonies in <100 mL of water, thus eliminating the need to collect sperm with pipettes. This enabled collection of  $10^9$  sperm per day (a 100-fold increase). We froze the sperm at a concentration of  $5 \times 10^7$ /mL at a cooling rate of  $20^\circ\text{C}/\text{min}$ . When thawed, these sperm samples remained at a concentration of  $5 \times 10^7$  sperm/mL (Table 20, Experiment 2; Figure 76). Moreover, the number of fertilized eggs increased to 10. Only DMSO was used in this experiment as we decided to focus on one cryoprotectant.



**Figure 76. Cooling curve for Experiment 2, 20°C/min.**

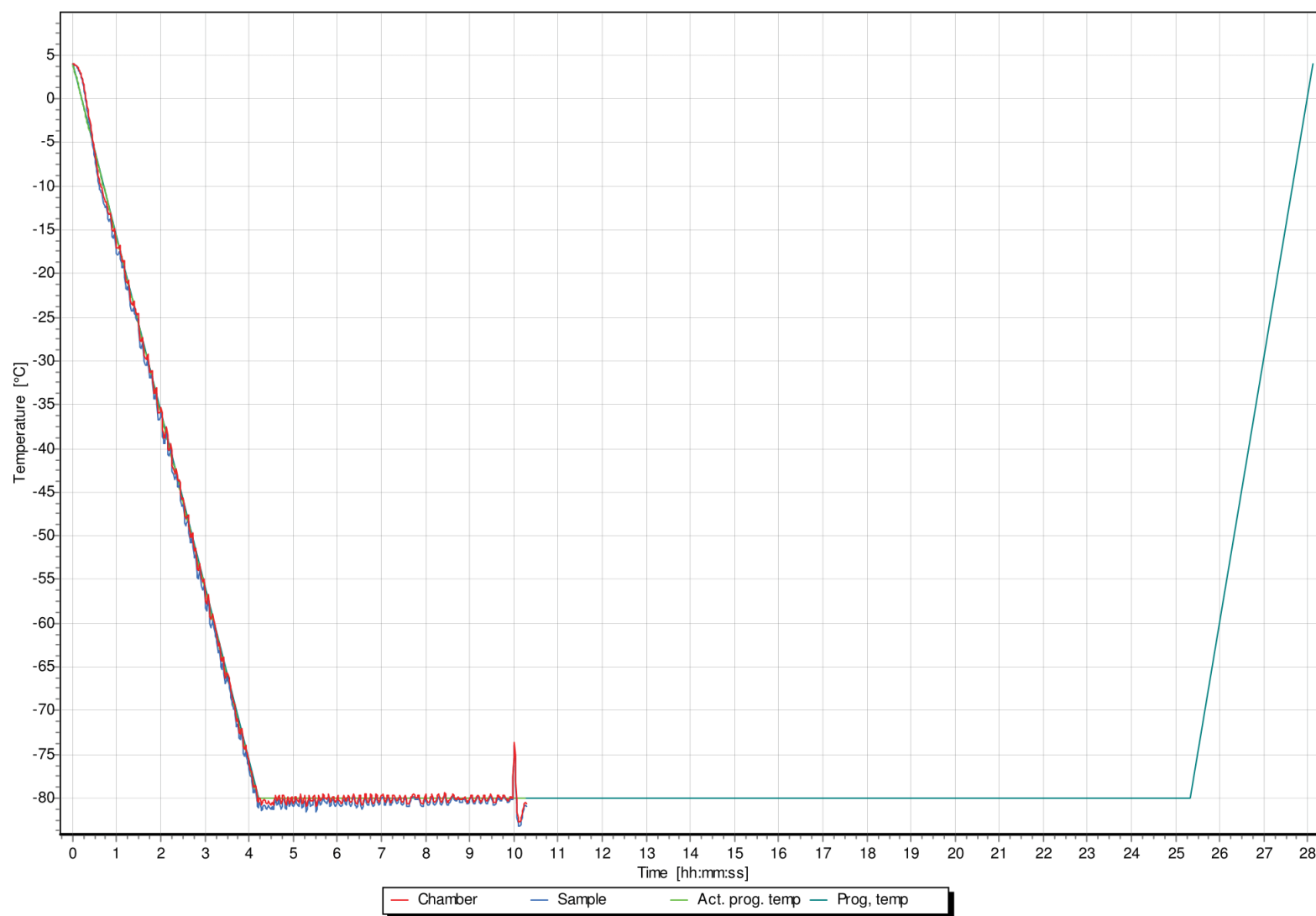
These results encouraged us to test whether we could further increase fertilization by increasing the concentration of sperm samples. We froze sperm at five different concentrations ranging from  $10^7$  to  $10^9$ /mL (Figure 77) at the  $20^\circ\text{C}/\text{min}$  cooling rate (Figure 78) and stored them in liquid nitrogen for 21 hr. The concentration of each sample post-thaw had the same count as before freezing. When thawed, sperm frozen at  $10^9$ /mL were able to fertilize 270-275 eggs. Sperm (in descending order of concentration) at  $5 \times 10^8$ /mL fertilized 235-250 eggs;  $1 \times 10^8$ /mL fertilized 150-250 eggs;  $5 \times 10^7$ /mL fertilized 26-150 eggs, and  $1 \times 10^7$ /mL fertilized 5-100 eggs. All embryos developed into larvae and were able to metamorphose into a primary polyp with no visual abnormalities. Thus, cooling sperm at a rate of  $20^\circ\text{C}/\text{min}$  and at concentrations in excess of  $1 \times 10^7$  showed best fertilization.



**Figure 77. Fertilization comparing frozen sperm.**

Each thawed sperm sample was exposed to different number of eggs. In each case, the number of eggs collected was manually estimated to be a surplus of what each respective sperm sample could fertilize based on their concentration.





**Figure 78. Cooling curve for variable sperm concentration at 20°C/min.**

## 7.5 Discussion

### 7.5.1 Sperm motility and viability

The sperm motility and concentration-related phenomena reported herein provide some insight of the basic characteristics of sperm clouds that have not been previously observed. While the results from this feasibility study are promising, there are several improvements and future experiments that can be pursued. There was a large standard deviation (<50%) in motility and concentration between individually collected sperm streams. Part of this variation reflects the imprecise manner traditionally used for collecting sperm as it is released. These findings reinforce the need to standardize collection methods and sperm concentrations. The results also suggest that it is possible to store sperm for at least 6d at 4°C without an appreciable drop in fertilization capability, thereby enabling shipment of sperm samples. This also demonstrated that sperm motility is not necessarily a good predictor of fertilization success when gametes are mixed under controlled conditions. Future studies can also address other outstanding questions related to these characteristics. For example, when and how are sperm activated? Does the sperm concentration affect activation and motility? Determination of how these features could affect cryopreservation, especially among different genotypes, would be useful in expanding and making protocols more robust for *Hydractinia* and potentially other cnidarian models.

Concepts such as these have been studied quantitatively in aquatic species previously at the commercial scale, for example in blue catfish, *Ictalurus furcatus*, (for hatchery production of hybrids) (Hu et al., 2014) by use of industrial engineering and simulation modeling approaches (Hu et al., 2015). Those studies were based on use of automated high-throughput processing (Hu et al., 2011) developed using commercial dairy industry approaches (Lang et al., 2003) but are also

relevant for processing at lower throughput. The emphasis in such approaches is on the application level for repository development, rather than on the research level for optimization of individual components (e.g., cooling rate or cryoprotectant choice) for protocol development.

Another application-level concept often overlooked in traditional research approaches is the refrigerated storage of samples prior to freezing or use. Such storage enables shipping of germplasm for processing elsewhere and can avoid waste by identifying the usable working lifetime of valuable material. We tested the retention of *Hydractinia* sperm fertility after storage in FSW at 4°C and found that freshly collected sperm and sperm stored for 6 d could fertilize comparable numbers of eggs. This result suggested that sperm could be stored in FSW at 4°C even longer and still produce viable embryos. Identifying these basic storage conditions is useful in cases when resources are not available to process on-site and samples must be transported to another facility for processing and storage.

Future studies should compare fertility across a range of storage temperatures with longer storage times when appropriate, and couple that with freezing experiments to evaluate the effects of storage on cryopreservation survival. In addition, extender solutions (e.g., buffers) can influence the quality and retention of fertility of sperm during storage (Paniagua-chavez et al., 2000; Tiersch, 2011b; H. Yang et al., 2018). Future studies should also address the total fertilization window for eggs. While mixing gametes  $\leq 30$  min post-release has been the community guideline for producing quality embryos, this has not been determined quantitatively and it is possible that storage at a cooler temperature may extend fertility.

### 7.5.2 Determining cryoprotectant toxicity to sperm

The acute toxicity assay we performed was at a small scale but yielded useful information regarding potential cryoprotectants. We initially observed limited fertilization using the treated sperm, which demonstrated feasibility and a basis for improvement. In future studies, the potential effects of cryoprotectant toxicity on sperm and egg should be evaluated more clearly. If toxicity is affecting fertilization, sperm can be rinsed to reduce or eliminate the cryoprotectant before exposing them to the eggs. The limited fertilization we observed also emphasized the need to process sperm in concentrations that were relevant to those used for breeding. This prompted the design of a custom 3-D printed collection chamber to improve sperm collection, and enabled evaluation of cryopreservation conditions that resulted in effective post-thaw fertilization rates. This improved collection method provides expanded opportunities for standardized evaluation of cryoprotectants and concentrations, while bearing in mind that such choices should be governed by overall utility at the process level rather than optimizing singular factors (e.g., motility) at the individual step level. For example, a certain cryoprotectant may yield a slightly lower motility value than other chemicals, but is cheaper, less toxic to sperm cells, and allows more flexibility in timing and cooling rates. In research-driven studies, the highest motility would be recommended; in application-driven work, the cryoprotectant that increases efficiency and reliability would be recommended.

Other benefits of placing a focus on application include that work in the present study can be directly scaled up for use with hundreds of animals and multiple laboratories. Work addressing repository development in previous studies, with blue catfish for example, can be generalized to *Hydractinia* because the approaches used are the same, including the use of French straws that can be filled, sealed, and labelled using automated equipment (e.g., the Minitube Quattro system at the

AGGRC can process 15,000 straws/hr). In addition, cryopreservation in *Hydractinia* can be directly transferred from a central facility (such as the AGGRC) to on-site work within an existing laboratory by use of high-throughput mobile cryopreservation capabilities (Childress et al., 2018), or by establishment of full high-throughput cryopreservation capabilities such as in creation of a central *Hydractinia* Stock Center (for economic analysis, see (Gwo, 2018)). Development of in-house cryopreservation capabilities within research laboratories will be greatly strengthened by the recent developments in 3-D printing described above (e.g., (Hu et al., 2017)) including fabrication of probes for monitoring and storing temperature information (Shamkhalichenar et al., 2019), and the potential for sharing of open-source design files for production of inexpensive, reproducible freezing devices that can be integrated with strong quality control programs (e.g., (Hu et al., 2013; Leticia Torres et al., 2016)).

### **7.5.3 Identifying suitable freezing conditions**

While there are no other *Hydractinia* cryopreservation protocols to directly compare our results to, there are protocols that have been developed for sperm from various coral species, which can serve as an indirect comparison for some of the key parameters. One protocol in particular has been instrumental in banking the germplasm of 31 coral species from around the world (Mary Hagedorn, Carter, et al., 2012; Mary Hagedorn, Oppen, et al., 2012; Mary Hagedorn et al., 2019) and additional protocols have been developed in two coral species (Ohki et al., 2014; Viyakarn et al., 2018). Briefly, we can compare our method with the cryoprotectant, container, and cooling methods of these studies. Similar to two of the studies (Mary Hagedorn, Carter, et al., 2012; Viyakarn et al., 2018), DMSO was used as the cryoprotectant but at higher final concentrations ( $\geq 10\%$ ), and in the other (Ohki et al., 2014) 20% methanol was used with 0.9 M sucrose as an

extender. These cryoprotectant concentrations are higher than what our trials suggested would be suitable for *Hydractinia* sperm. However, there are two major differences that may explain this discrepancy and proffer improvements to this study. The *in vitro* fertilization in this study used a volumetric sperm to egg ratio of 50 mL with  $\sim 10^5$ /mL sperm to 30-40 eggs. This was considerably more dilute in comparison to each coral study in which the ratios used to determine post-cryoprotectant fertility were 5 mL with  $10^6$ /mL sperm to 30-50 eggs (Mary Hagedorn, Carter, et al., 2012), 1 mL with  $10^5$ /mL sperm to 20 eggs (Ohki et al., 2014), or 4 mL with  $1.5 \times 10^7$ /mL sperm to 50 eggs (Viyakarn et al., 2018). In future acute toxicity assays, optimizing the volumetric sperm to egg ratio (in our case, reducing the volume and increasing the sperm concentration) would improve the assessment of acute toxicity before moving onto freezing. Previous studies with eastern oyster *Crassostrea gigas* have shown that much of the variation in sperm cryopreservation response is procedural rather than biological (e.g., “male-to-male variation”) and control of sperm concentration is necessary for reproducible results (Dong et al., 2007).

One of the coral protocols cryopreserved 1-mL samples in 2-mL cryovials (Mary Hagedorn, Carter, et al., 2012), whereas the other two studies (Ohki et al., 2014; Viyakarn et al., 2018) cryopreserved samples in 0.25-mL French straws. French straws offer several advantages over traditional cryovials. French straws require less storage space and can be easily processed manually in the case of a few samples, or more efficiently in high-throughput with automated filling, labeling, and sealing for hundreds to thousands of samples. In addition, samples can generally be cooled in French straws at a faster rate than in cryovials, in large part due to the higher surface-area-to-volume ratio of straws, which can also decrease variability during freezing. In cryovials, there is potentially more variation across the sample volume as material on the periphery

could freeze more rapidly than that closer to the center. Also, vials typically have thicker walls with greater insulative potential slowing heat removal from the sample.

With regard to cooling rates, there are several differences that make these studies difficult to compare. First, the equilibration temperature and time used were slightly different (Mary Hagedorn, Carter, et al., 2012; Viyakarn et al., 2018) or not explicitly quantified (Ohki et al., 2014). Second, the ending temperatures used to calculate the freezing curve were different where one study used  $-80^{\circ}\text{C}$  (Mary Hagedorn, Carter, et al., 2012), but the other two used the coldest achievable temperature between  $-110^{\circ}\text{C}$  and  $-130^{\circ}\text{C}$ . Theoretically, the ending temperature should not affect the rate calculation if the freezing rate is constant, but unless the temperature is monitored while the samples are being frozen, fluctuations are difficult to account for. Although the different procedures make studies difficult to compare, it is critical that all details surrounding the freezing process be documented for quality samples and reproducible results (Torres & Tiersch, 2018). For this reason, only two of the studies can be referenced for reproducibility and generally compared in relation to their cooling rate (Mary Hagedorn, Carter, et al., 2012; Viyakarn et al., 2018). Both studies used an equilibration temperature between  $24\text{--}29^{\circ}\text{C}$  and equilibration time of 15 min (Viyakarn et al., 2018) or 20 min (Mary Hagedorn, Carter, et al., 2012) whereas in this study the equilibration temperature was  $4^{\circ}\text{C}$  for 20 min. The selection of  $4^{\circ}\text{C}$  as the equilibration temperature in our study was in part due to the usage of a controlled-rate freezer.

One obvious difference among the present study and the three published coral studies is that each used suspension at defined heights above liquid nitrogen to freeze samples. This method is difficult to standardize, and is less precise than using a controlled-rate freezer. However, it has significant advantages, such as affordability, availability, and portability. In future studies, a comparison of samples frozen at comparable nominal rates by various methods should be done to

enable harmonization of results and reporting, providing multiple options for freezing that could be selected based on the user's needs. Other factors that could be investigated in future work include whether offspring produced from cryopreserved sperm will mature into full adults and whether the male-to-female ratio is affected.

#### **7.5.4 Approaches to repository development for aquatic species**

Recent advances in consumer-level technology provide opportunities to custom-design open-source options for hardware and other tools necessary to assist repository development beyond that provided by adaptation of traditional livestock practices. Customizing the design of the 3-D printed collection chamber greatly increased the efficiency and success in identifying suitable freezing conditions. To collect a useful number of sperm, the standard collection method via Pasteur pipette or micropipette is labor intensive and poses logistical problems in the case of multiple collectors. Given our previous approach, collecting all sperm would be possible but would require filtering all the water from the bin (~2 L) or having access to a large capacity centrifuge. Thus, by customizing a chamber to minimize the collection volume (<100 mL) and maximize the total yield of sperm (as many as ten slides bearing *Hydractinia*), we were able to directly improve and standardize processing efficiency.

In addition, custom design of devices is also possible for freezing activities. The polylactic acid (PLA) used for 3-D printing does not become brittle or stiff as do other plastics when exposed to cryogenic temperatures (such as liquid nitrogen) (Tiersch & Monroe, 2016), making 3-D printed objects safe and useful for such applications. Various devices can typically be fabricated at low cost (e.g., \$10 or less for material costs) using consumer-level printers (\$250 or less) that offer high resolution, flexibility, and short learning curves. There are large internet-driven user



communities for these printers and thousands of videos (such as on YouTube) are available for printer set up, training, and troubleshooting. In addition, design files can be shared on a number of sites (e.g., Thingiverse and Github) for others to print and customize. In this way, devices used in cryopreservation and repository development can be developed, shared, and standardized within research communities, greatly reducing costs of cryopreservation, and making reliable methods widely available. Systems such as this must be accompanied by quality control and quality assurance programs, however, to ensure that samples meet minimum thresholds for repository use (Hu et al., 2013; Leticia Torres et al., 2016).)

Overall, the success in the present study of using a generalizable approach for *Hydractinia* sperm provides further evidence that cryopreservation protocols are not necessarily species-specific. For example, a single generalized protocol was applied to more than 20 species within the genus *Xiphophorus* and two other species in the genus *Poecilia* to enable repository development to safeguard the genetic resources of these valuable biomedical model species (Liu, Torres, et al., 2018). Overall, more research is needed for aquatic species in general to quantitatively assess factors important to practical repository operation with cryopreserved sperm (e.g., (Hu et al., 2014)), and standardization of procedures and reporting is necessary to enable meaningful comparisons across studies (Torres & Tiersch, 2018). The present study offers evidence that substantial repository-level benefits can be realized by generalizing cryopreservation at the application level, rather than trying to optimize new protocols on a species-by-species basis, and restricting this work to the traditional (reductionist) research level (Torres & Tiersch, 2018).

## 7.6 Conclusions

This feasibility study showed that it is possible and a worthwhile endeavor to pursue *Hydractinia* sperm cryopreservation as a long-term storage option for genetic resources. Specifically, we demonstrated that sperm cooled at 20°C per min in 5% DMSO at a concentration of  $10^8$ - $10^9$ /ml in 0.25-mL French straws were able to fertilize 150-300 eggs, which developed into juvenile colonies. In our experience, a population of 150 juvenile colonies typically contains sufficient numbers to establish a strain for propagation via asexual reproduction (i.e., they will grow into healthy adults with the genotype of interest) or breeding to produce subsequent generations. With some additional work, it should be possible to reliably freeze and re-derive specific genotypes. This would greatly enhance the utility of *Hydractinia* as a model system for cnidarian genetics. Establishing repository capabilities for the *Hydractinia* research community will be essential for future development, maintenance, protection, and distribution of genetic resources. More broadly, this application-based approach highlights the long-term value of establishing repository-level resources that can be expanded to fit community needs. In addition, we expect this work could also provide a guide to researchers seeking to develop cryopreservation approaches in other cnidarian species.

While this study has direct implications for the *Hydractinia* community, there are several considerations that can be discussed with regard to communities that work on other cnidarian models. Lack of long-term storage options has been one of the limitations to nearly all cnidarian research. Cryopreservation has not been pursued either due to the lack of resources to achieve and maintain frozen samples, or the lack of necessity as many cnidarian models can be cultured relatively simply and the animals can regenerate. A notable exception to this are cryopreservation efforts for conservation in coral species and their symbionts due to importance of corals to reef

biodiversity, and the overall decline in health and prevalence of corals globally over the past several decades (Hoegh-Guldberg et al., 2018, 2019; Palumbi et al., 2019; Weis, 2019).

Another emerging problem for popular research models is the rapid proliferation of new lines and mutants that would require maintenance as live animals, which is expensive and risky without cryopreservation. With these common limitations in mind, cnidarian communities need to come together and agree on a consistent and foundational approach towards cryopreservation of all cnidarian models for the ultimate purpose of repository development and establishment of repository networks. By having this long-term goal in mind, we can more systematically work towards developing, protecting, maintaining, distributing, and utilizing an expanding pool of cnidarian genetic resources.

A centralized repository or stock center is a necessity for well-developed research organisms. Part of the success with these repositories can be attributed to collaboration among laboratories and the sharing of tools, systems, and resources throughout the communities. For example, mouse resources are largely centralized with The Jackson Laboratory (<https://www.jax.org/>), zebrafish databases and lines are found in the Zebrafish International Resources Center (ZIRC, University of Oregon), *Drosophila* utilizes the Bloomington *Drosophila* Stock Center (BDSC, Indiana University Bloomington), *Caenorhabditis elegans* and other worm-related models localize their resources in WormBase ([wormbase.org/](http://wormbase.org/)), and *Xenopus* related resources are found in Xenbase (<https://www.xenbase.org>). Having a wealth of such resources and information available to these communities makes these model systems much more useful and available to investigators, whereas model systems that require development of basic tools can be more challenging on many levels.

Future studies should establish a standardized approach for the storage, shipment, and use of frozen *Hydractinia* samples that can be made available throughout the research community. Current models for this would include development of repositories or a repository system, and the potential incorporation of these entities into a community-based stock center. An existing model for such organization exists in ZIRC, which maintains more than 43,000 research lines of zebrafish as frozen sperm (<https://zebrafish.org/>). In addition, to assist standardization of protocols and approaches, it may be useful to establish community-level mechanisms to design and share inexpensive devices that can be used to support users across a wide range of experience and skill levels in culture, spawning, and cryopreservation of *Hydractinia*. Lastly, cryopreservation and repository development should be expanded to include additional germplasm and somatic cell types.

## **7.7 Acknowledgements**

We thank Mallory Lemoine, William Childress, Amy Guitreau, Liu Yue, Teresa Gutierrez-Wing of the AGGRC for technical assistance and discussion. This manuscript was approved for publication by the Director of the Louisiana Agricultural Experiment Station as number 2021-241-34993.

## **7.8 Materials and methods**

### **7.8.1 Ethics**

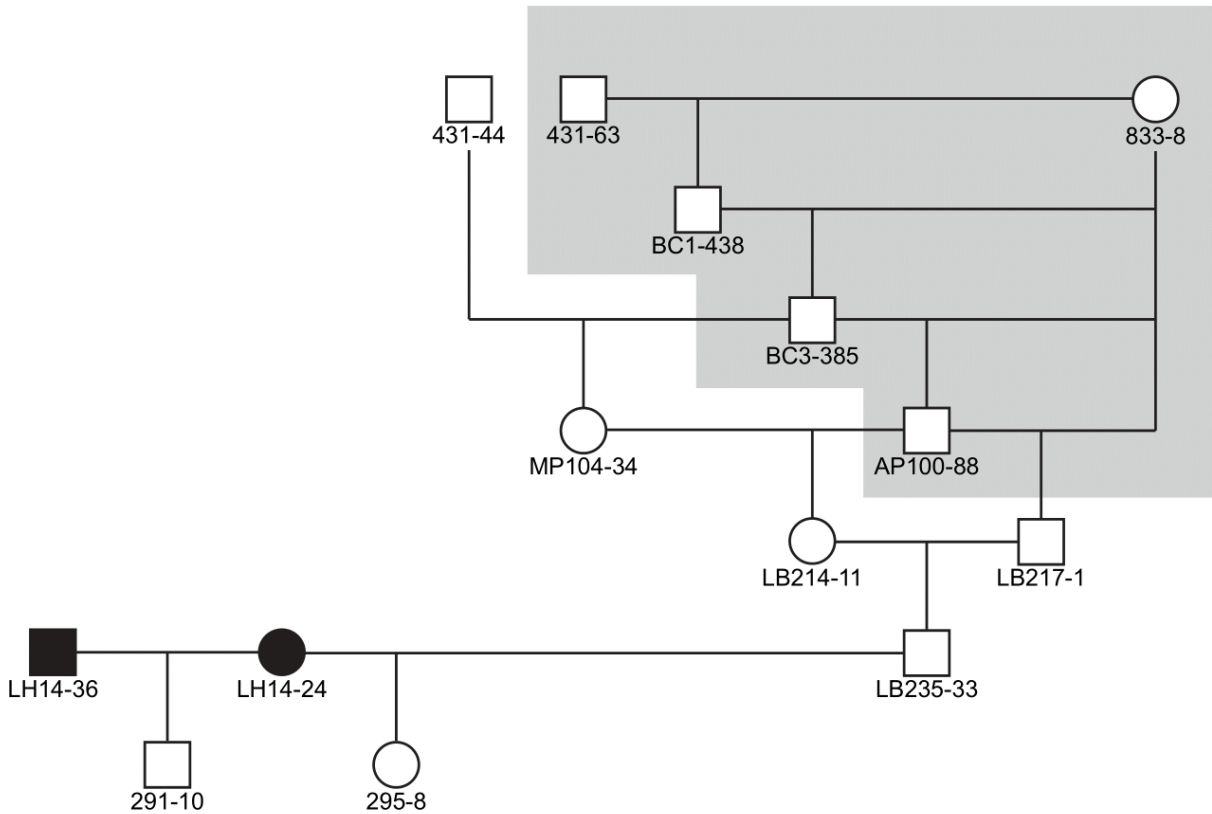
Animal care is overseen by separate Institutional Animal Care and Use Committees at the University of Pittsburgh and Louisiana State University. *Hydractinia symbiolongicarpus* is a marine invertebrate lacking a central nervous system and is not regulated by specialized guidelines. All animals used in this study were maintained in continuous culture as detailed below.

### **7.8.2 Animal care and breeding**

Experimental work was performed from February to April 2019, at the Aquatic Germplasm & Genetic Resources Center (AGGRC) in Baton Rouge. Animals were transported in 50-mL tubes by overnight shipping from the University of Pittsburgh. Colonies were maintained and grown as previously described (Sanders et al., 2018) and cultured for 2 weeks before experimental use. Briefly, colonies were established on 25 mm x 75 mm glass microscope slides and cultured in 38-L (10-gal) aquaria using artificial seawater (ASW) (Instant Ocean Reef Crystals, Spectrum Brands, Blacksburg, VA) at between 29 and 31 ppt, held at 22-23°C, and maintained on an 8h:16h (light:dark) photoperiod. Adult colonies were fed 4-day-old *Artemia* nauplii three times per week at 48-hr intervals. Twice at 24-hr between *Artemia* feedings, colonies were fed a suspension of pureed oysters. Oysters were freshly caught, shucked, pureed, aliquoted, flash frozen in liquid nitrogen, and stored at -20°C.

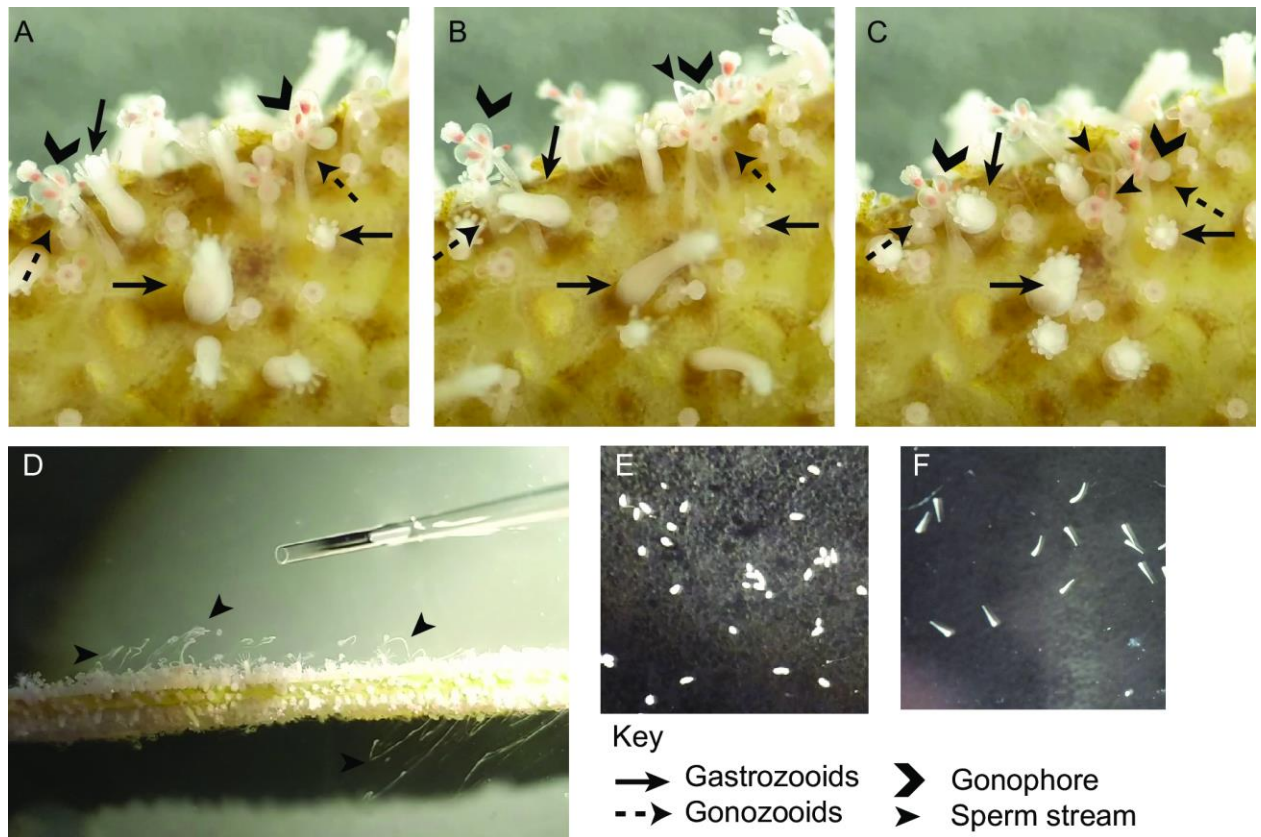
In this study, we performed crosses between two half siblings, a male (colony 291-10) and a female (colony 295-8) (Figure 79). Following first exposure to light, male and female colonies

were moved into separate bins filled with ASW and placed under supplemental lighting. Gametes released approximately 1 hr after light exposure (Ballard, 1942). Sperm were released in “clouds” or “streams” from individual gonophores (Figure 80A-C) and were collected and pooled using a Pasteur pipette (Figure 80D).



**Figure 79. Pedigree of the colonies used to generate germplasm and offspring.**

Field-collected colonies are denoted with black symbols. Colony 291-10 is the offspring of two colonies collected from Lighthouse Point, New Haven, CT in 2014. Colony 295-8 is the offspring of a field collected colony and a laboratory strain, 235-33. The pedigree of colony 235-33 can be recreated by concatenating previously published pedigrees (shaded area) (Cadavid et al., 2004; Powell et al., 2007). Colony AP100-88 is from the mapping population in Powell et al. (2007). Colony 431-44 is from the mapping population in Cadavid et al. (2004).



**Figure 80. Time lapse of sperm release.**

(A) Close-up view of *Hydractinia* polyps just prior to sperm release. Arrows indicate polyp types. (B) Arrowhead points to sperm stream being released. (C) Arrowheads point to sperm stream (polyps have retracted from B). (D) Top-down view of slide with *Hydractinia* releasing sperm. Arrowheads point to multiple streams of sperm released from the colony. (E) 1-d old larvae. (F) 2-d old larvae.

Eggs were collected by straining the water from the female bin with a 20- $\mu$ m cell strainer. For routine breeding and to serve as a positive control for fertilization, 20-30 clouds of sperm were collected from 10 male slides, transferred to a 50-mL conical tube, and brought to a final volume of 15 mL with filtered sea water (FSW, artificial seawater filtered through 0.45  $\mu$ m Polyethersulfone (PES) membrane Rapid-Flow Sterile Disposable Bottle Top Filters, Thermo Scientific Nalgene, catalog #295-4545). To this were added 400-600 eggs harvested from 8-9 female slides. The final volume was brought to 30 mL with FSW and transferred to a 100-mm

polystyrene Petri dish. Within 1 hr, embryos began to cleave and developed into planulae by the following day (Figure 80E). On day 4 after fertilization, larvae (Figure 80F) were settled by exposure to 100 mM Cesium Chloride (CsCl diluted in FSW) for 4-5 hr until ready for settlement, and were pipetted onto microscope slides and kept in the dark for 1-2 d or until attachment and primary polyps formed.

### **7.8.3 Estimation of sperm concentration and motility**

On six separate days, individual sperm clouds (cumulative N = 35) were collected in a 10- $\mu$ l volume and analyzed for motility within 20 min of collection. The sample was briefly vortexed to form a uniform suspension, loaded onto a Makler<sup>®</sup> counting chamber (SEFI Medical Instruments Ltd, Irvine Scientific, Santa Ana, CA, USA), and viewed with dark-field illumination at 200-X magnification (Olympus CX41RF, Tokyo, Japan). Sperm were already motile when observed and did not require activation. The sample concentration was counted twice according to an established protocol (Liu, Torres, et al., 2018) and the average was used as the sperm concentration (at 10<sup>6</sup>/mL). Motility was quantified using a computer-assisted sperm analysis (CASA) system (CEROS model; Hamilton Thorne, Inc., Beverly, MA, USA) which was set up at room temperature. Any samples that were incubated at different temperatures were quantified as quickly as possible to achieve a reading at their incubated temperature. The settings used were based on a previous study (Liu, Yang, et al., 2018). Briefly, motility and VCL (curvilinear velocity) were measured for 10 sec over the whole sample. Cell detection was set at a minimum of 25 pixels for contrast and 6 pixels for cell size. In each individual sperm measurement, 100 frames were captured at a rate of 60 frames/s. Sperm with an average of >20  $\mu$ m/s measured path velocity (VAP) were counted by the program as being progressively motile. GraphPad Prism (v8.2.0) was



used to calculate correlations between sperm characteristics (velocity, percent motile, and concentration).

#### **7.8.4 Longevity and temperature sensitivity of sperm**

To test the effects of time and temperature on sperm motility, approximately 30 sperm clouds were collected using a 10- $\mu$ L pipette, pooled and diluted to produce a concentration of  $2 \times 10^7$  cells/mL (approximately 0.7 mL total), and then divided into two tubes. One tube was kept at room temperature (21-23°C unless otherwise stated) and the other was kept in a 4°C refrigerator. Each treatment was evaluated hourly for presence or absence of motility for the first 7 hr and then visualized again at 23 hr for presence or absence of motility.

#### **7.8.5 Fertility of sperm**

To determine how long sperm could produce viable offspring when stored at 4°C, we performed a time-series experiment using a single collection of sperm. Approximately 150 clouds of sperm were collected using a Pasteur pipette and stored in a 50-mL conical tube. Concentration was determined as described above. On day 0, 3 mL of this sample (total of  $2 \times 10^7$  sperm) were used to fertilize 200 eggs in a total volume of 30 mL FSW. The sperm sample was stored at 4°C. On subsequent days (up to Day 6), freshly collected eggs were fertilized with 3 mL ( $2 \times 10^7$  sperm) of sperm in 30 mL FSW. Offspring from each day's fertilization experiment were followed until they metamorphosed into juvenile polyps.

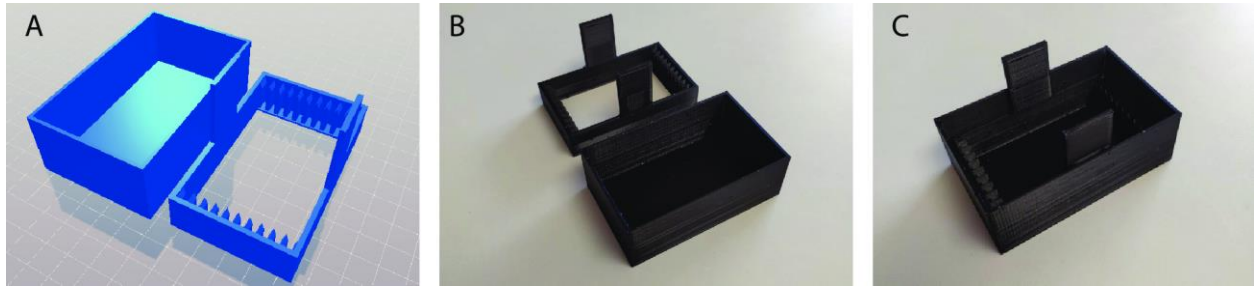
### 7.8.6 Acute Toxicity of Cryoprotectants

Approximately 20 sperm clouds were collected using a 10- $\mu$ L pipette, pooled, and adjusted to a concentration of  $1 \times 10^7$  sperm/mL using FSW. Three cryoprotectants, methanol (Fisher Scientific, Waltham, MA) dimethyl sulfoxide (DMSO, Fisher Scientific, Waltham, MA), and glycerol (Sigma-Aldrich, St. Louis, MO) were used. For each cryoprotectant, double strength stocks of 10%, 20%, and 30% (v/v) were created using FSW. The sperm and double-strength cryoprotectant were mixed in equal volumes (100  $\mu$ L:100  $\mu$ L) at room temperature resulting in a final sperm concentration of  $5 \times 10^6$  sperm/mL and final cryoprotectant concentrations of 5%, 10%, or 15%. Evaluating cryoprotectant concentrations less than 5% was not necessary as *Hydractinia* sperm were not excessively toxic. Likewise, concentrations greater than 15% were not tested as 15% was already too toxic. Sperm were evaluated at 30 min after addition of cryoprotectant at room temperature, which was a practical total exposure time required for cryoprotectant equilibration and for packaging and handling of the samples. Presence or absence of motility was used as an estimate for toxicity.

### 7.8.7 Standardized Sperm Collection (3-D printing)

Based on the difficulties and inefficiencies experienced during pilot experiments working with *Hydractinia* sperm, we designed a custom sperm collection chamber with integrated slide rack to collect and concentrate sperm for downstream applications (Figure 81) by use of free computer-aided design (CAD) online software (Tinkercad, version 4.7, Autodesk, San Rafael, CA). The design was exported as a stereolithography (STL) file and imported into a 3-D printer slicer software (Simplify3D, version 4.0, Cincinnati, OH) to control the printing process (Table

21) Collection chambers were printed in black PLA (ZYLtech Engineering, Spring, TX) filament on a stock Prusa i3 MK3 3-D printer (Prusa Research, Czech Republic) (Table 22).



**Figure 81. Sperm collection chamber.**

(A) CAD-rendering of the 3-D design. (B) Printed model with rack and box separate. (C) Printed model with rack inserted. Object model deposited on Thingiverse. <https://www.thingiverse.com/thing:3661286>

**Table 21. Slicer software settings used to 3-D print collection chamber.**

<b>Settings</b>	<b>Expression</b>
<b>General settings</b>	
Hotend temperature	200°C
Print speed	60 mm/s
Nozzle type	Brass
Nozzle diameter	0.4 mm
Extrusion/line width	0.45 mm
Nominal layer height	0.2 mm
Retraction distance	1.5 mm
Retraction speed	30 mm/s
Printer bed temperature	60 °C
Part cooling fan speed	75%
<b>First layer settings</b>	
Extrusion/line width	0.45 mm
Layer height	0.2 mm
Print speed	36 mm/s
Heat block temperature	205 °C
<b>Part specific settings</b>	
Infill	100%
Infill pattern	Rectangular
Wall/perimeter layers	2
Top layers	3
Bottom layers	3
Support placement	None
Support overhang angle	n/a
Support density	n/a
Build surface adhesion	Skirt

**Table 22. Printer hardware features.**

<b>Variable</b>	<b>Expression</b>
Printer	Prusa i3 MK3
Power supply voltage	24 v
Extrusion	Direct drive
Filament size	1.75 mm
Filament supplier and type	PLA
Filament storage conditions	63-L plastic bin
Build surface size	21 x 21 cm
Build surface	PEI
Cooling fan	Blower
Cooling fan size	51.5 x 51.5 x 15 mm
Cooling fan voltage	5 v
Auto bed leveling sensor	PINDA proximity sensor

### **7.8.8 Freezing**

To collect sperm for freezing, we placed nine slides of males in the 3-D printed sperm collection chamber filled with ASW. An additional male was placed in a separate bin so that sperm could be collected and used as a fertilization positive control. After sperm were released, the slide rack was removed from the sperm collection chamber and the cloudy seawater was poured into two 50-mL conical tubes (~80 mL total) and spun for 20 min at 3,000 rpm (~1450-1500 x g) at room temperature which resulted in a visible white pellet. The supernatant was pipetted off and the pellets were combined and resuspended in FSW to the appropriate concentrations (between  $2 \times 10^6$  and  $2 \times 10^9$  sperm/mL) and stored at 4°C until they were prepared for freezing (~3 hr).

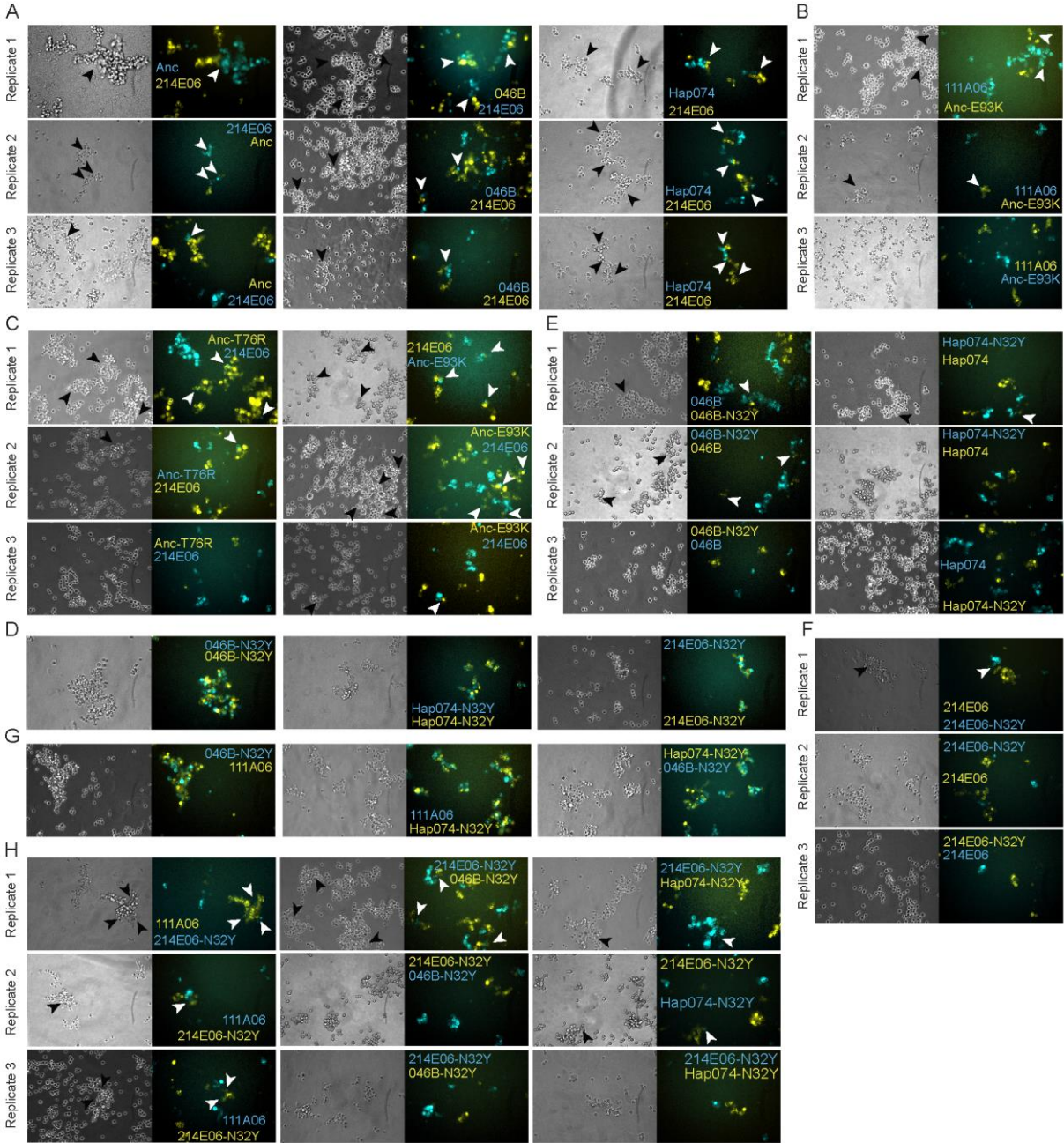
To prepare for freezing, sperm were mixed with an equal volume of 10% DMSO or 10% methanol in FSW (final concentrations of 5% cryoprotectant), drawn into 0.25-mL French straws (IMV International, MN, USA), and held at 4°C in a controlled-rate freezer for the remaining equilibration time (Minitube of America, IceCube 14M, SY-LAB). The total equilibration time,

from initial mixing with cryoprotectant to starting the freezing program, was set at 20 min. Equilibrated samples were cooled to  $-80^{\circ}\text{C}$  with one of three pre-programmed cooling rates:  $5^{\circ}\text{C}/\text{min}$ ,  $20^{\circ}\text{C}/\text{min}$ , or  $30^{\circ}\text{C}/\text{min}$ . Frozen samples were held at  $-80^{\circ}\text{C}$  for at least 5 min before transfer and storage in liquid nitrogen.

#### **7.8.9 Thawing and use for fertilization**

After 21-69 hr of storage, straws were removed from liquid nitrogen and immediately plunged into room temperature ( $22^{\circ}\text{C}$ ) water for 8 s. The straws were clipped and a  $2\text{-}\mu\text{L}$  sample was removed, diluted with  $38\text{ }\mu\text{L}$  of FSW (1:20 dilution), and used for sperm assessment. The remaining sample was held in a microfuge tube until fresh eggs were obtained (15-30 min). After performing the fertilization positive control (routine breeding), 100-300 fresh eggs were collected in  $500\text{ }\mu\text{L}$  of FSW and added to the microfuge tube with the thawed sperm. The mixed gametes were placed into a 100-mm Petri dish and  $\sim 50\text{ mL}$  FSW was added. An estimate of the number of eggs used was obtained by counting in groups of ten. The resulting fertilization was kept at room temperature and observed for 24 hr to determine how many planulae had begun forming. The resulting offspring were observed until metamorphosis into juvenile polyps.

Appendix A Supplemental Figures

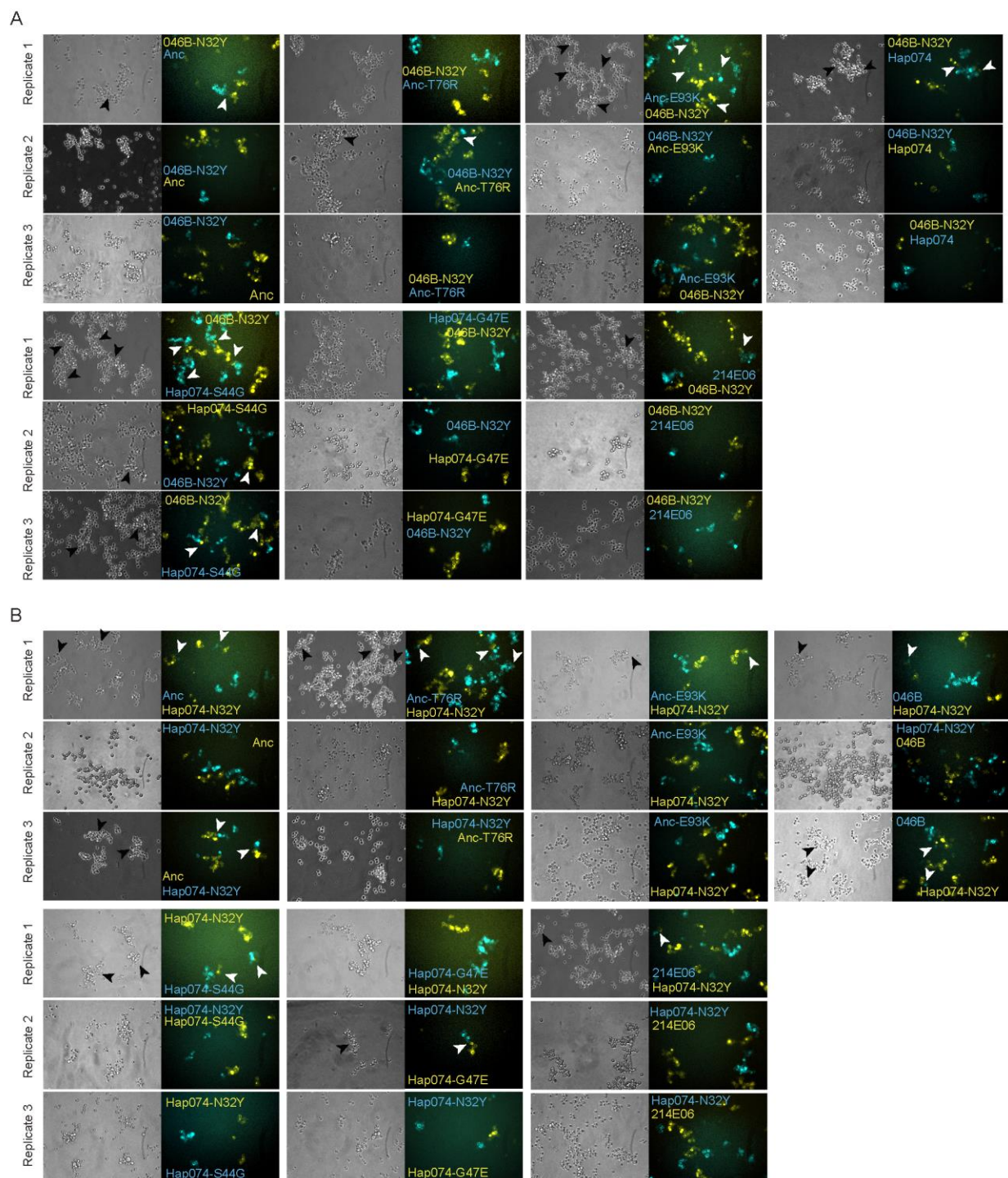


Supplemental Figure 1. Replicate cell aggregation assay results related to Figure 46, Figure 48, Figure 49, Figure 50, Figure 51, Figure 52.

Caption on following page.

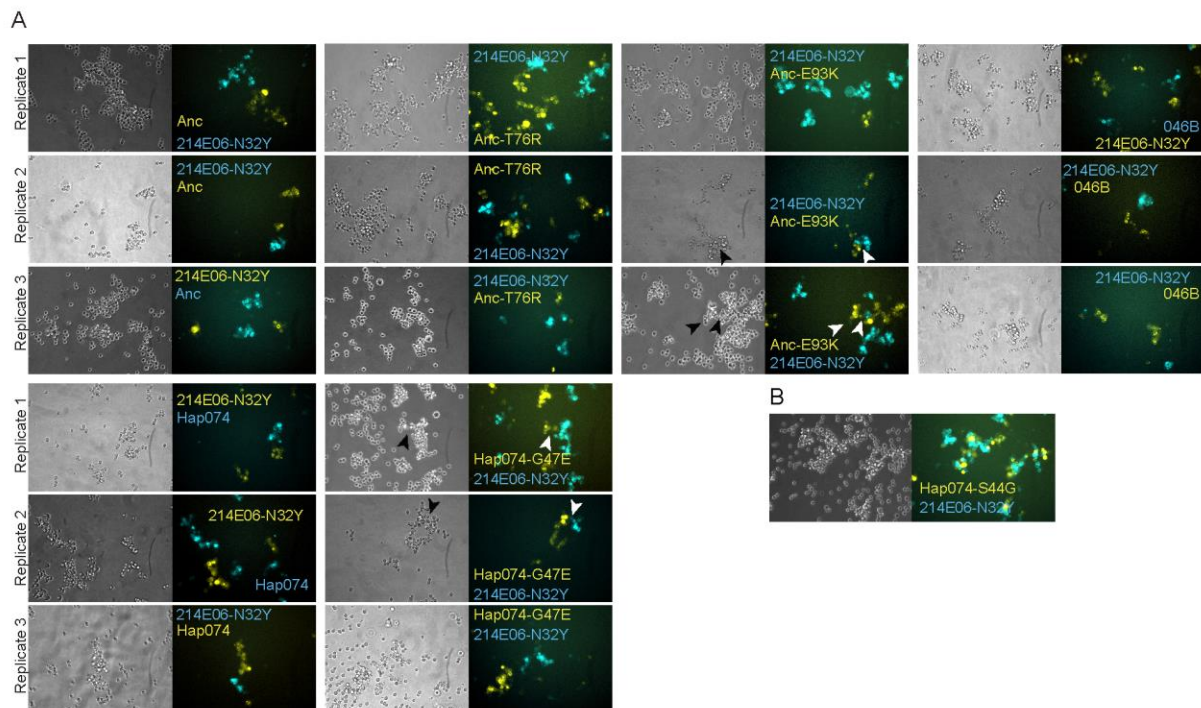
Supplemental Figure 1 caption: Arrowheads indicate possible semi-mixed aggregates. A) Replicates of Figure 46 consistently yield semi-mixed aggregates indicating either non-specific interactions or weak heterophilic interactions. B) Replicates of Figure 49B result in semi-mixed aggregates in N=2 replicates. C) Replicates of Figure 49C result in semi-mixed aggregates in N=2 replicates for Anc-T76R vs 214E06 and N=3 replicates for Anc-E93K vs 214E06. D) Homophilic binding results for N32Y mutants. E) Both 046B-N32Y and Hap074-N32Y form semi-mixed aggregates with their immediate ancestor in at least one replicate. F) 214E06-N32Y forms mixed aggregates with 214E06 in one replicate. G) Both 046B-N32Y and Hap074-N32Y bind homophilically to 111A06 and to each other. F) 214E06-N32Y does not form mixed aggregates with 111A06, 046B-N32Y, or Hap074-N32Y though does result in some semi-mixed aggregates.





**Supplemental Figure 2. Replicate cell aggregation assay results related to Figure 52.**

Arrowheads indicate possible semi-mixed aggregates. A) Replicates of aggregation assay results with 046B-N32Y yield semi-mixed aggregates in most cases. B) Replicates of aggregation assay results with Hap074-N32Y yield semi-mixed aggregates in most cases.



**Supplemental Figure 3. Replicate cell aggregation assay results related to Figure 52.**

Arrowheads indicate possible semi-mixed aggregates. A) Replicates of aggregation assay results with 214E06-N32Y yield semi-mixed aggregates in most cases. B) 214E06-N32Y vs Hap074-S44G yield mixed aggregates indicating homophilic binding between the pair.

## Appendix B Supplemental Tables

**Supplemental Table 1. Contigs encoding *Alr* genes in 291-10 assembly and their classification.**

Classification is based on the coloring scheme as before (bona fide in blue, putative in orange, pseudogenes in black). Contig names correspond to their assembly. Primary contigs in the ARC-se haplotype are due to the 11% duplication found by BUSCO in the assembly after the filtering process. ARC-se (2) contains contigs on which a third haplotype was identified but cannot be resolved in its position in the ARC.

	<i>pr</i>	<i>se</i>	<i>se2</i>
Alr1	HyS0031	hsym_2_tig00003089_polished	HyS2844
Alr2	HyS0001	hsym_2_tig00000211_polished	
Alr3	HyS0029	hsym_2_tig00001187_polished	
Alr4	HyS0029	hsym_2_tig00009368_polished	
Alr5/Alr5p	HyS0029	hsym_2_tig00001791_polished	hsym_2_tig00001685_polished
Alr6	HyS0029	HyS0131	
Alr7	HyS0029	HyS0131	
Alr8p	HyS0029	HyS0131	
Alr9	HyS0029	HyS0131	
Alr9B	HyS0029		
Alr10p	HyS0029		
Alr11	HyS0029	HyS0131	hsym_2_tig00002752_polished
Alr11p		HyS0131	
Alr12			hsym_2_tig00002752_polished
Alr12A	HyS0029	HyS0131	hsym_2_tig00002199_polished
Alr12B	HyS0029	HyS0131	
Alr12C	HyS0029		
Alr12Dp		HyS0131	
Alr12Ep		HyS0131	
Alr14	HyS0029	HyS0131	hsym_2_tig00002199_polished
Alr15	HyS0031		
Alr15p		hsym_2_tig00009001_polished	hsym_2_tig00002106_polished
Alr16/Alr16p	HyS0031		hsym_2_tig00002106_polished
Alr17	HyS0031	hsym_2_tig00003089_polished	
Alr18	HyS0031		
Alr18A	HyS0031		
Alr19	HyS4647		
Alr19A		hsym_2_tig00003089_polished	
Alr19B		hsym_2_tig00003089_polished	
Alr20/Alr20p	HyS0037	hsym_2_tig00003089_polished	
Alr20A	HyS0031		

Supplemental Table 1 continued on next page

Supplemental Table 1 continued

Alr20B	HyS0031	
Alr20C	HyS0031	
Alr21	HyS0031	hsym_2_tig00003089_polished
Alr22	HyS0031	hsym_2_tig00009001_polished
Alr23		hsym_2_tig00009001_polished
Alr24p	HyS0031	hsym_2_tig00009001_polished
Alr25p	HyS0022	
Alr26p	HyS0022	
Alr27	HyS0022	hsym_2_tig00003071_polished
Alr28	HyS0022	hsym_2_tig00001753_polished
Alr29	HyS0022	hsym_2_tig00001753_polished
Alr30		hsym_2_tig00000211_polished
Alr30A	HyS0001	
Alr30B	HyS0001	
Alr31	HyS0001	hsym_2_tig00000211_polished
Alr32	HyS0001	hsym_2_tig00000211_polished
Alr33	HyS0001	hsym_2_tig00000204_polished
Alr34	HyS0001	hsym_2_tig00000229_polished
Alr35	HyS0001	hsym_2_tig00000229_polished
Alr36/Alr36p	HyS0001	hsym_2_tig00000253_polished
Alr37	HyS0001	hsym_2_tig00000542_polished
Alr38	HyS0001	hsym_2_tig00000600_polished
Alr38p	HyS0031	

---

**Supplemental Table 2. *Alr* gene classification comparison between the F and 291-10 haplotypes.**

Classification is based on the coloring scheme as before (bona fide in blue and putative in orange).

All annotations	<i>F</i>	<i>pr</i>	<i>se</i>	<i>se2</i>
<a href="#">Alr1</a>	<a href="#">Alr1</a>	<a href="#">Alr1</a>	<a href="#">Alr1</a>	<a href="#">Alr1</a>
<a href="#">Alr2</a>	<a href="#">Alr2</a>	<a href="#">Alr2</a>	<a href="#">Alr2</a>	
Alr2p2 <sup>a</sup>	Alr2p2		( <a href="#">Alr30B</a> ) <sup>a</sup>	
<a href="#">Alr3</a>	<a href="#">Alr3</a>	<a href="#">Alr3</a>	<a href="#">Alr3</a>	
<a href="#">Alr4</a>	<a href="#">Alr4</a>	<a href="#">Alr4</a>	<a href="#">Alr4</a>	
Alr5		<a href="#">Alr5</a>	<a href="#">Alr5</a>	
Alr5p	Alr5p		Alr5p	
<a href="#">Alr6</a>	<a href="#">Alr6</a>	<a href="#">Alr6</a>	<a href="#">Alr6</a>	
Alr7	<a href="#">Alr7</a>	<a href="#">Alr7</a>	<a href="#">Alr7</a>	
Alr8	<a href="#">Alr8</a>			
Alr8p		Alr8p	Alr8p	
Alr9	<a href="#">Alr9</a>	<a href="#">Alr9</a>	<a href="#">Alr9</a>	
Alr9B			Alr9B	
Alr10				
Alr10Ap			Alr10Ap	
Alr10p	Alr10p			
Alr11	<a href="#">Alr11</a>	<a href="#">Alr11</a>	<a href="#">Alr11</a>	<a href="#">Alr11</a>
Alr11p		Alr11p		
Alr12			<a href="#">Alr12</a>	
<a href="#">Alr12A</a>	<a href="#">Alr12A</a>	<a href="#">Alr12A</a>	<a href="#">Alr12A</a>	<a href="#">Alr12A</a>
<a href="#">Alr12B</a>	<a href="#">Alr12B</a>	<a href="#">Alr12B</a>	<a href="#">Alr12B</a>	
Alr12C		<a href="#">Alr12C</a>		
Alr12Cp	Alr12Cp			
Alr12Dp		Alr12Dp		
Alr12Ep		Alr12Ep	Alr12Ep	
Alr13p	Alr13p			
Alr14		<a href="#">Alr14</a>	<a href="#">Alr14</a>	<a href="#">Alr14</a>
Alr14p	Alr14p			
Alr15	<a href="#">Alr15</a>	<a href="#">Alr15</a>	Alr15p	
Alr15p				15p
Alr16	<a href="#">Alr16</a>	<a href="#">Alr16</a>	Alr16p	
Alr17	<a href="#">Alr17</a>	<a href="#">Alr17</a>	<a href="#">Alr17</a>	
Alr18	<a href="#">Alr18</a>	<a href="#">Alr18</a>		
Alr18A		<a href="#">Alr18A</a>		
Alr19	<a href="#">Alr19</a>	<a href="#">Alr19</a>		
Alr19A			<a href="#">Alr19A</a>	
Alr19B			<a href="#">Alr19B</a>	
Alr20		<a href="#">Alr20</a>		
Alr20A		<a href="#">Alr20A</a>		

Supplemental Table 2 continued on next page

Supplemental Table 2 continued

Alr20B		Alr20B	
Alr20C		Alr20C	
Alr20Dp			Alr20Dp
Alr20p	Alr20p		
Alr21	Alr21	Alr21	Alr21
Alr22		Alr22	Alr22
Alr22p	Alr22p		
Alr23	Alr23	Alr23	Alr23
Alr24p	Alr24p	Alr24p	Alr24p
Alr25p	Alr25p	Alr25p	
Alr26p	Alr26p	Alr26p	
Alr27	Alr27	Alr27	Alr27
Alr28	Alr28	Alr28	Alr28
Alr29	Alr29	Alr29	Alr29
Alr30	Alr30		Alr30
Alr30A		Alr30A	
Alr30B <sup>a</sup>	(Alr2p2) <sup>a</sup>	Alr30B	
Alr31	Alr31	Alr31	Alr31
Alr32		Alr32	Alr32
Alr32p	Alr32p		
Alr33	Alr33	Alr33	Alr33
Alr34	Alr34	Alr34	Alr34
Alr35	Alr35	Alr35	Alr35
Alr36	Alr36	Alr36	Alr36p
Alr37	Alr37	Alr37	Alr37
Alr38	Alr38	Alr38	Alr38
Alr38p		Alr38p	

<sup>a</sup> The equivalent allele for Alr2p2 was renamed Alr30B to avoid confusion with Alr2.



**Supplemental Table 3. *Alr* ectodomain variation compared between haplotypes (extended table).**

Gene	Polymorphic sites		Length	% ID			Isoforms compared <sup>c</sup>
	(%) <sup>a</sup>	(#) <sup>b</sup>		F vs. wt1	F vs. wt2	wt1 vs. wt2	
Alr1	62.69%	210	335-353	80.24	37.43	34.72	3
Alr2	44.77%	184	398-411	68.24	61.98	79.4	3
Alr3	2.35%	8	341	97.95	99.71	97.65	3
Alr4	1.48%	5	337	98.81	99.41	99.81	3
Alr5	0%	0	208			100	2
Alr5p	8.13%	10	123-168		90.32		2
Alr6	34.29%	108	314-315	72.15	81.27	69.94	3
Alr7	3.64%	12	330	96.97	97.58	98.18	3
Alr9	6.75%	22	326	100	93.25	93.25	3
Alr11	12.84%	42	327	98.16	88.07	88.69	3
Alr12B	19.20%	62	311-323	83.64	92.73	97.94	3
Alr15	7.96%	25	304-314	88.85			2
Alr16	1.75%	6	342	100	98.25	98.25	3
Alr17	3.60%	5	139	97.12	98.56	97.12	3
Alr18	14.16%	47	332		85.84		2
Alr21	3.68%	13	353	96.64	96.64	98.47	3
Alr22	1.54%	5	324	98.46	98.46	100	3
Alr23	5.81%	20	226-344	92.48	92.48	100	3
Alr27	9.71%	30	309	100	90.29	90.29	3
Alr28	2.52%	8	317	98.74	97.79	98.42	3
Alr29	0.90%	2	222	99.55	99.1	99.55	3
Alr30	14.78%	60	404-406		85.75		2
Alr31	1.80%	6	333	98.05	99.67	98.37	3
Alr32	5.84%	15	255-257	90.2	94.9	94.96	3
Alr33	4.00%	13	325	97.54	97.85	96.62	3
Alr34	1.23%	4	325	100	98.77	98.77	3
Alr35	1.26%	4	318	100	98.74	98.74	3
Alr36	2.88%	13	451	98.11	90.87	90.7	3
Alr37	2.46%	5	203	98.52	98.03	98.52	3
Alr38	0.56%	2	356	99.39	99.69	99.69	3

<sup>a</sup> The shorter length is used for calculation. Genes with >25% polymorphic sites are highlighted in green.

<sup>b</sup> Gaps are not counted as polymorphic sites.

<sup>c</sup> Certain genes are not present in all three haplotypes. See Supplemental Table 2 for these genes.

**Supplemental Table 4. *Alr* cytoplasmic tail variation compared between haplotypes (extended table).**

Gene	Polymorphic sites		Length	% ID			Isoforms compared <sup>c</sup>
	(%) <sup>a</sup>	(#) <sup>b</sup>		F vs. wt1	F vs. wt2	wt1 vs. wt2	
Alr1	9.03%	14	155	97.42	92.26	92.26	3
Alr2	8.60%	19	221	91.82	91.82	98.19	3
Alr3	3.37%	6	178	96.63	96.63	100	3
Alr4	1.77%	2	113	99.12	99.12	98.23	3
Alr5	2.67%	2	75			97.33	2
Alr6	4.24%	5	118	99.15	95.76	96.61	3
Alr7	1.41%	1	69-71	98.55	97.1	98.55	3
Alr9	9.59%	7	73	91.78	90.41	98.63	3
Alr9B	19.18%	14	73	86.3	84.93	83.56	1
Alr11	8.22%	6	73	97.26	91.78	94.52	3
Alr12A	4.11%	3	73	97.26	97.26	97.26	3
Alr12B	5.48%	4	73	95.89			2
Alr14 <sup>d</sup>	15.49%	11	71	84.51	98.59	85.92	3
Alr15 <sup>e</sup>	0.00%	0	89	100	100	100	3
Alr16	0.00%	0	45-78	100	100	100	3
Alr17	0.00%	0	10	100	100	100	3
Alr18	12.35%	10	81	87.65			2
Alr19	1.28%	1	78	98.72			2
Alr21	2.50%	2	80	97.50	97.50	100	3
Alr22 <sup>f</sup>	7.41%	6	81	92.59	92.59	100	3
Alr23	7.84%	4	51	92.16	92.16	100	3
Alr27	0.00%	0	47	100	100	100	3
Alr28	0.89%	1	112	99.11	100	99.11	3
Alr29	1.83%	2	109	98.17	100	98.17	3
Alr30	2.44%	2	82		97.56		2
Alr31	0.93%	1	107	99.07	100	99.07	3
Alr32 <sup>g</sup>	6.09%	7	115	93.91	100	93.91	2
Alr33	7.27%	4	55	96.36	96.36	92.73	3
Alr34	3.38%	10	296	94.26	93.92	99.66	3
Alr36	1.33%	1	75	98.67			2
Alr37	0.00%	0	37	100	100	100	3
Alr38	6.25%	3	46-48	93.75	100	93.75	3

<sup>a</sup> The shorter length is used for calculation. Genes with >5% polymorphic sites are highlighted in green.

<sup>b</sup> Gaps are not included in as part of the number of polymorphic sites

<sup>c</sup> Certain genes are not present in all three haplotypes. See Supplemental Table 2 for these genes.

<sup>d</sup> Alr14p-F is compared with Alr14-wt1 and Alr14-wt2 cytoplasmic tails.

<sup>e</sup> Alr15p-F is compared with Alr15-wt1 and Alr15-wt2 cytoplasmic tails.

<sup>f</sup> Alr22p-F is compared with Alr22-wt1 and Alr22-wt2 cytoplasmic tails.

<sup>g</sup> Alr32p-F is compared with Alr32-wt1 and Alr32-wt2 cytoplasmic tails.



## Bibliography

- Aanen, D. K., Debets, A. J. M., Visser, J. A. G. M. De, & Hoekstra, R. F. (2008). The social evolution of somatic fusion. *BioEssays*, 30(11–12), 1193–1203. <https://doi.org/10.1002/bies.20840>
- Abràmoff, M. D., Magalhães, P. J., & Ram, S. J. (2004). Image Processing with ImageJ. *Biophotonics International*, 11(7), 36–42.
- Afgan, E., Baker, D., Beek, M. Van Den, Bouvier, D., Chilton, J., Clements, D., Coraor, N., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46, 537–544. <https://doi.org/10.1093/nar/gky379>
- Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., Heijne, G. von, & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, 37, 420–423. <https://doi.org/10.1038/s41587-019-0036-z>
- Au-yeung, B. B., Shah, N. H., Shen, L., & Weiss, A. (2017). ZAP-70 in Signaling, Biology, and Disease. *Annual Review of Immunology*, 36, 109–138.
- Ballard, W. W. (1942). The Mechanism for Synchronous Spawning in Hydractinia and Pennaria. *Biological Bulletin*, 82(3), 329–339.
- Barrow, A. D., & Trowsdale, J. (2006). You say ITAM and I say ITIM, let's call the whole thing off: the ambiguity of immunoreceptor signalling. *European Journal of Immunology*, 36, 1646–1653. <https://doi.org/10.1002/eji.200636195>
- Bastiaans, E., Debets, A. J. M., & Aanen, D. K. (2015). Experimental demonstration of the benefits of somatic fusion and the consequences for allorecognition. *Evolution*, 69(4), 1091–1099. <https://doi.org/10.1111/evo.12626>
- Bastiaans, E., Debets, A. J. M., Aanen, D. K., Diepeningen, A. D. Van, Saupe, S. J., & Paoletti, M. (2014). Natural Variation of Heterokaryon Incompatibility Gene het-c in *Podospira anserina* Reveals Diversifying Selection. *Molecular Biology and Evolution*, 31(4), 962–974. <https://doi.org/10.1093/molbev/msu047>
- Beadle, G. W., & Coonradt, V. L. (1944). Heterocaryosis in *Neurospora crassa*. *Genetics*, 29, 291–308.
- Benabentos, R., Hirose, S., Sugang, R., Curk, T., Katoh, M., Ostrowski, E., Strassmann, J., Queller, D., Zupan, B., Shaulsky, G., & Kuspa, A. (2009). Polymorphic members of the lag-

- gene family mediate kin-discrimination in *Dictyostelium*. *Current Biology*, 19(7), 567–572. <https://doi.org/10.1016/j.cub.2009.02.037>. Polymorphic
- Benisty, S., Ben-Jacob, E., Ariel, G., & Be'er, A. (2015). Antibiotic-Induced Anomalous Statistics of Collective Bacterial Swarming. *Physical Review Letters*, 114(018105), 1–5. <https://doi.org/10.1103/PhysRevLett.114.018105>
- Berlin, K., Koren, S., Chin, C., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6), 623–630. <https://doi.org/10.1038/nbt.3238>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.
- Bernet, J. (1967). Les systèmes d'incompatibilité chez le *Podospira anserina*. *Comptes Rendus de l'Académie Des Sciences*, 265, 1330–1333.
- Bezbradica, J. S., & Medzhitov, R. (2012). Role of ITAM signaling module in signal integration. *Current Opinion in Immunology*, 24, 58–66. <https://doi.org/10.1016/j.coi.2011.12.010>
- Blackstone, N. W. (1996). Gastrovascular Flow and Colony Development in Two Colonial Hydroids. *Biological Bulletin*, 190, 56–68.
- Blakeslee, A. F. (1904). Zygosporangium Formation a Sexual Process. *Science*, 19(492), 864–866. <https://doi.org/10.1126/science.19.492.864>
- Bod'ova, K., Priklopil, T., Field, D. L., Barton, N. H., & Pickup, M. (2018). Evolutionary Pathways for the Generation of New Self-Incompatibility Haplotypes in a Nonself-Recognition System. *Genetics*, 209(July), 861–883.
- Bodelon, G., Palomino, C., & Fernandez, L. A. (2013). Immunoglobulin domains in *Escherichia coli* and other enterobacteria: from pathogenesis to applications in antibody technologies. *Federation of European Microbiological Societies*, 37, 204–250. <https://doi.org/10.1111/j.1574-6976.2012.00347.x>
- Bradshaw, B., Thompson, K., & Frank, U. (2015). Distinct mechanisms underlie oral vs aboral regeneration in the cnidarian *Hydractinia echinata*. *ELife*, 4(e05506), 1–19. <https://doi.org/10.7554/eLife.05506>
- Buckley, W. J., & Ebersole, J. P. (1994). Symbiotic organisms increase the vulnerability of a hermit crab to predation. *Journal of Experimental Marine Biology and Ecology*, 182(1), 49–64.
- Buss, L. W. (1982). Somatic cell parasitism and the evolution of somatic tissue compatibility. *Proceedings of the National Academy of Science*, 79(September), 5337–5341.
- Buss, L. W. (1987). *The Evolution of Individuality*. Princeton University Press.

- Buss, L. W., Anderson, C., Westerman, E., Kritzberger, C., Poudyal, M., Moreno, M. A., & Lakkis, F. G. (2012). Allorecognition Triggers Autophagy and Subsequent Necrosis in the Cnidarian *Hydractinia symbiolongicarpus*. *PLoS ONE*, 7(11), 1–10. <https://doi.org/10.1371/journal.pone.0048914>
- Buss, L. W., McFadden, C. S., & Kenne, D. R. (1984). Biology of Hydractiniid Hydroids. 2. Histocompatibility Effector System/Competitive Mechanism Mediated By Nematocyst Discharge. *Biological Bulletin*, 167(August), 139–158. <https://doi.org/10.2307/1541430>
- Cadavid, L. F., Powell, A. E., Nicotra, M. L., Moreno, M., & Buss, L. W. (2004). An invertebrate histocompatibility complex. *Genetics*, 167(1), 357–365. <https://doi.org/10.1534/genetics.167.1.357>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 9(421), 1–9. <https://doi.org/10.1186/1471-2105-10-421>
- Cannon, J. P., Haire, R. N., & Litman, G. W. (2002). Identification of diversified genes that contain immunoglobulin-like variable regions in a protochordate. *Nature Immunology*, 3(12), 1200–1207. <https://doi.org/10.1038/ni849>
- Cao, P., & Wall, D. (2017). Self-identity reprogrammed by a single residue switch in a cell surface receptor of a social bacterium. *Proceedings of the National Academy of Sciences*, 114(14), 3732–3737. <https://doi.org/10.1073/pnas.1700315114>
- Cao, P., & Wall, D. (2019). Direct visualization of a molecular handshake that governs kin recognition and tissue formation in myxobacteria. *Nature Communications*, 10(3073), 1–10. <https://doi.org/10.1038/s41467-019-11108-w>
- Cao, P., Wei, X., Awal, P., & Müller, R. (2019). A Highly Polymorphic Receptor Governs Many Distinct Self-Recognition Types within the Myxococcales Order. *MBio*, 10(1), 1–15.
- Cardarelli, L., Saak, C., & Gibbs, K. A. (2015). Two Proteins Form a Heteromeric Bacterial Self-Recognition Complex in Which Variable Subdomains Determine Allele-Restricted Binding. *MBio*, 6(3), 1–8. <https://doi.org/10.1128/mBio.00251-15>
- Casselton, L. A., & Olesnick, N. S. (1998). Molecular Genetics of Mating Recognition in Basidiomycete Fungi. *Microbiology and Molecular Biology Reviews*, 62(1), 55–70.
- Caten, C. E. (1972). Vegetative Incompatibility and Cytoplasmic Infection in Fungi. *Journal of General Microbiology*, 72, 221–229.
- Catherine S. McFadden, M. J. M., & Buss, L. W. (1984). Biology of Hydractiniid Hydroids . 1 . Colony Ontogeny in *Hydractinia echinata* (Flemming). *Biological Bulletin*, 166(1), 54–67.
- Chan, T. A., Chu, C. A., Rauen, K. A., Kroihner, M., Tatarewicz, S. M., & Steele, R. E. (1994). Identification of a gene encoding a novel protein-tyrosine kinase containing SH2 domains and ankyrin-like repeats. *Oncogene*, 9(4), 1253–1259.

- Chandonia, J.-M., Fox, N. K., & Brenner, S. E. (2019). SCOPe: classification of large macromolecular structures in the structural classification of proteins — extended database. *Nucleic Acids Research*, 47, 475–481. <https://doi.org/10.1093/nar/gky1134>
- Chang, S., & Staben, C. (1994). Directed Replacement of mt A by mt a-1 Effects a Mating Type Switch in *Neurospora crassa*. *Genetics*, 81, 75–81.
- Charlesworth, D. (2006). Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, 2(4), 2–7. <https://doi.org/10.1371/journal.pgen.0020064>
- Chen, G., Wang, J., Xu, X., Wu, X., Piao, R., & Siu, C.-H. (2013). TgrC1 mediates cell–cell adhesion by interacting with TgrB1 via mutual IPT/TIG domains during development of *Dictyostelium discoideum*. *Biochemical Journal*, 452(2), 259–269. <https://doi.org/10.1042/BJ20121674>
- Chen, G., Xu, X., Wu, X., Thomson, A., & Siu, C.-H. (2014). Assembly of the TgrB1–TgrC1 cell adhesion complex during *Dictyostelium discoideum* development. *Biochemical Journal*, 459(2), 241–249. <https://doi.org/10.1042/BJ20131594>
- Childress, W. M., Caffey, R. H., & Tiersch, T. R. (2018). Design and Cost Analysis of a Self-contained Mobile Laboratory for Commercial-scale Aquatic Species Cryopreservation. *World Aquaculture Society*, 49(5), 805–826. <https://doi.org/10.1111/jwas.12525>
- Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6), 563–569. <https://doi.org/10.1038/nmeth.2474>
- Chittor, A., & Gibbs, K. A. (2021). The conserved serine transporter SdaC moonlights to enable self recognition. *BioRxiv*, 1–49.
- Chookajorn, T., Kachroo, A., Ripoll, D. R., Clark, A. G., & Nasrallah, J. B. (2004). Specificity determinants and diversification of the Brassica self-incompatibility pollen ligand. *Proceedings of the National Academy of Sciences*, 101(4), 911–917.
- Chou, K.-C. (2000). Prediction of Tight Turns and Their Types in Proteins. *Analytical Biochemistry*, 286, 1–16. <https://doi.org/10.1006/abio.2000.4757>
- Colombatti, A., Bonald, P., & Doliana, R. (1993). Type A Modules: Interacting Domains Found in Several Non-Fibrillar Collagens and in Other Extracellular Matrix Proteins. *Matrix*, 13(4), 297–306.
- Copeland, M. F., & Weibel, D. B. (2009). Bacterial Swarming: A Model System for Studying Dynamic Self-assembly. *Soft Matter*, 5(6), 1174–1187. <https://doi.org/10.1039/B812146J>
- Coppin, E., Debuchy, R., Arnaise, S., & Picard, M. (1997). Mating Types and Sexual Development in Filamentous Ascomycetes. *Microbiology and Molecular Biology Reviews*, 61(4), 411–428.

- Cortesi, P., & Milgroom, M. G. (1998). Genetics of Vegetative Incompatibility in *Cryphonectria parasitica*. *Applied and Environmental Microbiology*, 64(8), 2988–2994.
- Cunningham, C. w, Buss, L. W., & Anderson, C. (1991). Molecular and Geologic Evidence of Shared History Between Hermit Crabs and the Symbiotic Genus *Hydractinia*. *International Journal of Organic Evolution*, 45(1990), 1301–1316.
- Czárán, T., Hoekstra, R. F., & Aanen, D. K. (2014). Selection against somatic parasitism can maintain allorecognition in fungi. *Fungal Genetics and Biology*, 73, 128–137. <https://doi.org/10.1016/j.fgb.2014.09.010>
- Dammer, U., Popescu, O., Wagner, P., Anselmetti, D., Guntherodt, H.-J., & Misevic, G. N. (1995). Binding Strength Between Cell Adhesion Proteoglycans Measured by Atomic Force Microscopy. *Science*, 267(5201), 1173–1175. <https://doi.org/10.1126/science.7855599>
- Daskalov, A., Gladieux, P., Heller, J., & Glass, N. L. (2019). Programmed Cell Death in *Neurospora crassa* Is Controlled by the Allorecognition Determinant *rcd-1*. *Genetics*, 213(December), 1387–1400.
- Daskalov, A., Heller, J., Herzog, S., Fleißner, A., & Glass, N. L. (2017). Molecular Mechanisms Regulating Cell Fusion and Heterokaryon Formation in Filamentous Fungi. *Microbiology Spectrum*, 5(2), 1–15. <https://doi.org/10.1128/microbiolspec.FUNK-0015-2016>
- De Tomaso, A. W. (2006). Allorecognition polymorphism versus parasitic stem cells. *Trends in Genetics*, 22(9), 485–490. <https://doi.org/10.1016/j.tig.2006.07.001>
- De Tomaso, A. W., Nyholm, S. V., Palmeri, K. J., Ishizuka, K. J., Ludington, W. B., Mitchel, K., & Weissman, I. L. (2005). Isolation and Characterization of a Protochordate Histocompatibility Locus. *Nature*, 438(7067), 454–459.
- De Tomaso, A. W., Saito, Y., Ishizuka, K. J., Palmeri, K. J., & Weissman, I. L. (1998). Mapping the Genome of a Model Protochordate. I. A Low Resolution Genetic Map Encompassing the Fusion/Histocompatibility (Fu/HC) Locus of *Botryllus schlosseri*. *Genetics*, 149, 277–287.
- De Tomaso, A. W., & Weissman, I. L. (2003). Initial characterization of a protochordate histocompatibility locus. *Immunogenetics*, 55, 480–490. <https://doi.org/10.1007/s00251-003-0612-7>
- Debets, A. J. M., & Griffiths, A. J. F. (1998). Polymorphism of het-genes prevents resource plundering in *Neurospora crassa*. *Mycology Research*, 102(11), 1343–1349.
- Delcher, A. L., Phillippy, A., Carlton, J., & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30(11), 2478–2483.
- Dey, A., Vassallo, C. N., Conklin, A. C., Pathak, D. T., Troselj, V., & Wall, D. (2016). Sibling Rivalry in *Myxococcus xanthus* Is Mediated by Kin Recognition and a Polyploid Prophage. *Journal of Bacteriology*, 198(6), 994–1004. <https://doi.org/10.1128/JB.00964-15>

- Dienes, L. (1946). Reproductive Processes in *Proteus* Cultures. *Experimental Biology and Medicine*, 62(2), 265–270. <https://doi.org/10.3181/00379727-63-15570>
- Dong, Q., Huang, C., & Tiersch, T. R. (2007). Control of sperm concentration is necessary for standardization of sperm cryopreservation in aquatic species: Evidence from sperm agglutination in oysters. *Cryobiology*, 54, 87–98. <https://doi.org/10.1016/j.cryobiol.2006.11.007>
- 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., Baxeavanis, A. D., & Frank, U. (2020). Transcription factor AP2 controls cnidarian germ cell induction. *Science*, 367(February), 757–762.
- Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., Rasche, H., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2019). Apollo: Democratizing genome annotation. *PLoS Computational Biology*, 15(2), 1–14.
- Dutheil, J., & Boussau, B. (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evolutionary Biology*, 8(255). <https://doi.org/10.1186/1471-2148-8-255>
- Dyer, P. S., Ingram, D. S., & Johnstone, K. (1992). THE CONTROL OF SEXUAL MORPHOGENESIS IN THE ASCOMYCOTINA. *Biological Reviews*, 67, 421–458.
- Dynes, J. L., Clark, A. M., Shaulsky, G., Kuspa, A., Loomis, W. F., & Firtel, R. A. (1994). LagC is required for cell-cell interactions that are essential for cell-type differentiation in *Dictyostelium*. *Genes & Development*, 8, 948–958.
- Dyrka, W., Lamacchia, M., Durrens, P., Kobe, B., Daskalov, A., Paoletti, M., Sherman, D. J., & Saupe, S. J. (2014). Diversity and Variability of NOD-Like Receptors in Fungi. *Genome Biology and Evolution*, 6(12), 3137–3158. <https://doi.org/10.1093/gbe/evu251>
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47, 427–432. <https://doi.org/10.1093/nar/gky995>
- English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., & Gibbs, R. A. (2012). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One*, 7(11), 1–12. <https://doi.org/10.1371/journal.pone.0047768>
- Entani, T., Iwano, M., Shiba, H., Che, F., Isogai, A., & Takayama, S. (2003). Comparative analysis of the self-incompatibility (S-) locus region of *Prunus mume*: identification of a pollen-expressed F-box gene with allelic diversity. *Genes to Cells*, 8, 203–213.
- Erickson, H. P. (2017). Protein unfolding under isometric tension — what force can integrins generate, and can it unfold FNIII domains? *Current Opinion in Structural Biology*, 42, 98–

105. <https://doi.org/10.1016/j.sbi.2016.12.002>

- Espagne, E., Balhade, P., Penin, M.-L., Barreau, C., & Turcq, B. (2002). HET-E and HET-D Belong to a New Subfamily of WD40 Proteins Involved in Vegetative Incompatibility Specificity in the Fungus *Podospira anserina*. *Genetics*, *161*, 71–81.
- Fernandez-Busquets, X., & Burger, M. M. (2003). Circular proteoglycans from sponges: first members of the spongican family. *Cellular and Molecular Life Sciences*, *60*, 88–112.
- Fernandez-Busquets, Xavier, & Burger, M. M. (1999). Cell Adhesion and Histocompatibility in Sponges. *Microscopy Research and Technique*, *44*, 204–218.
- Ferreira, A. V.-B., An, Z., Metzenberg, R. L., & Glass, N. L. (1998). Characterization of mat A-2, mat A-3 and matA Mating-Type Mutants of *Neurospora crassa*. *Genetics*, *148*, 1069–1079. <https://doi.org/10.1093/genetics/148.3.1069>
- Fox, N. K., Brenner, S. E., & Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins — extended , integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, *42*, 304–309. <https://doi.org/10.1093/nar/gkt1240>
- Frishman, D., & Argos, P. (1995). Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics*, *23*, 566–579.
- Fujii, S., Kubo, K., & Takayama, S. (2016). Non-self- and self-recognition models in plant self-incompatibility. *Nature Plants*, *2*(September), 1–9. <https://doi.org/10.1038/NPLANTS.2016.130>
- Fujimori, T., Nakajima, A., Shimada, N., & Sawai, S. (2019). Tissue self-organization based on collective cell migration by contact activation of locomotion and chemotaxis. *Proceedings of the National Academy of Science*, *116*(10), 1–6. <https://doi.org/10.1073/pnas.1815063116>
- Futosi, K., & Mócsai, A. (2016). Tyrosine kinase signaling pathways in neutrophils. *Immunological Reviews*, *273*, 121–139. <https://doi.org/10.1111/imr.12455>
- Gahan, J. M., Bradshaw, B., Flici, H., & Frank, U. (2016). The interstitial stem cells in *Hydractinia* and their role in regeneration. *Current Opinion in Genetics & Development*, *40*, 65–73. <https://doi.org/10.1016/j.gde.2016.06.006>
- Gaino, E., Bavestrello, G., & Magnino, G. (1999). Self/non-self recognition in sponges. *Italian Journal of Zoology*, *66*, 299–315. <https://doi.org/10.1080/11250009909356270>
- Gibbs, K. A., Urbanowski, M. L., & Greenberg, E. P. (2008). Genetic determinants of self identity and social recognition in bacteria. *Science (New York, N.Y.)*, *321*(5886), 256–259. <https://doi.org/10.1126/science.1160033>
- Gibbs, K. A., Wenren, L. M., & Greenberg, E. P. (2011). Identity Gene Expression in *Proteus mirabilis*. *Journal of Bacteriology*, *193*(13), 3286–3292. <https://doi.org/10.1128/JB.01167-10>

- Glass, N. L., & Dementhon, K. (2006). Non-self recognition and programmed cell death in filamentous fungi. *Current Opinion in Microbiology*, 9, 553–558. <https://doi.org/10.1016/j.mib.2006.09.001>
- Glass, N. L., Grotelueschen, J., & Metzenberg, R. L. (1990). *Neurospora crassa* A mating-type region. *Proceedings of the National Academy of Science*, 87(July), 4912–4916.
- Glass, N. L., & Kaneko, I. (2003). Fatal Attraction: Nonspecific Recognition and Heterokaryon Incompatibility in Filamentous Fungi. *Eukaryotic Cell*, 2(1), 1–8. <https://doi.org/10.1128/EC.2.1.1-8.2003>
- Glass, N. L., & Lee, L. (1992). Isolation of *Neurospora crassa* A Mating Type Mutants by Repeat Induced Point (RIP) Mutation. *Genetics*, 133, 125–133.
- Glass, N. L., Vollmer, S. J., Staben, C., Grotelueschen, J., Metzenberg, R. L., & Yanofsky, C. (1988). DNAs of the Two Mating-Type Alleles of *Neurospora crassa* Are Highly Dissimilar. *Science*, 241, 570–573.
- Gloria-Soria, A., Moreno, M. A., Yund, P. O., Lakkis, F. G., Dellaporta, S. L., & Buss, L. W. (2012). Evolutionary genetics of the hydroid allodeterminant *alr2*. *Molecular Biology and Evolution*, 29(12), 3921–3932. <https://doi.org/10.1093/molbev/mss197>
- Gonçalves, A. P., & Glass, N. L. (2020). Fungal social barriers: to fuse, or not to fuse, that is the question. *Communicative & Integrative Biology*, 13(1), 39–42. <https://doi.org/10.1080/19420889.2020.1740554>
- Goncalves, A. P., Heller, J., Span, E. A., Rosenfield, G., Do, H. P., Palma-Guerrero, J., Requena, N., Marletta, M. A., & 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. *Current Biology*, 29, 3006–3017. <https://doi.org/10.1016/j.cub.2019.07.060>
- Goodman, K. M., Yamagata, M., Jin, X., Manneppalli, S., Katsamba, P. S., Ahlsen, G., Sergeeva, A. P., Honig, B., Sanes, J. R., & Shapiro, L. (2016). Molecular basis of sidekick-mediated cell-cell adhesion and specificity. *eLife*, 5(September2016), 1–21. <https://doi.org/10.7554/eLife.19058>
- Grice, L. F., Gauthier, M. E. A., Roper, K. E., Fernandez-Busquets, X., Degnan, S. M., & Degnan, B. M. (2017). Origin and Evolution of the Sponge Aggregation Factor Gene Family. *Molecular Biology and Evolution*, 34(5), 1083–1099. <https://doi.org/10.1093/molbev/msx058>
- Grosberg, R. K. (1988). The Evolution of Allorecognition Specificity in Clonal Invertebrates. *The Quarterly Review of Biology*, 63(4), 377–412.
- Grosberg, R. K., & Quinn, J. F. (1986). The genetic control and consequences of kin recognition by the larvae of a colonial marine invertebrate. *Nature*, 332, 456–459.



- Gwo, J. (2018). *Cryopreservation of aquatic invertebrate semen : A review Cryopreservation of aquatic invertebrate semen : a review. March 2000.* <https://doi.org/10.1046/j.1365-2109.2000.00462.x>
- Hagedorn, M., Carter, V. L., Steyn, R. A., Krupp, D., Leong, J. C., Lang, R. P., & Tiersch, T. R. (2006). Preliminary studies of sperm cryopreservation in the mushroom coral, *Fungia scutaria*. *Cryobiology*, 52, 454–458. <https://doi.org/10.1016/j.cryobiol.2006.03.001>
- Hagedorn, Mary, Carter, V., Martorana, K., Paresa, M. K., Acker, J., Baums, I. B., Borneman, E., Brittsan, M., Byers, M., Henley, M., Laterveer, M., Leong, J., Mccarthy, M., Meyers, S., Nelson, B. D., Petersen, D., & Tiersch, T. (2012). Preserving and Using Germplasm and Dissociated Embryonic Cells for Conserving Caribbean and Pacific Coral. *PLoS ONE*, 7(3). <https://doi.org/10.1371/journal.pone.0033354>
- Hagedorn, Mary, Farrell, A., & Carter, V. L. (2013). Cryobiology of coral fragments. *Cryobiology*, 66(1), 17–23. <https://doi.org/10.1016/j.cryobiol.2012.10.003>
- Hagedorn, Mary, Oppen, M. J. H. Van, Carter, V., Henley, M., Abrego, D., Puill-stephan, E., Negri, A., Heyward, A., Macfarlane, D., & Spindler, R. (2012). First frozen repository for the Great Barrier Reef coral created. *Cryobiology*, 65(2), 157–158. <https://doi.org/10.1016/j.cryobiol.2012.05.008>
- Hagedorn, Mary, Spindler, R., & Daly, J. (2019). Cryopreservation as a Tool for Reef Restoration: 2019. In P. Comizzoli, J. L. Brown, & W. V Holt (Eds.), *Reproductive Sciences in Animal Conservation* (pp. 489–505). Springer International Publishing. [https://doi.org/10.1007/978-3-030-23633-5\\_16](https://doi.org/10.1007/978-3-030-23633-5_16)
- Halaby, D. M., Poupon, A., & Mornon, J.-P. (1999). The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Engineering*, 12(7), 563–571.
- Hall, C., Welch, J., Kowbel, D. J., & Glass, N. L. (2010). Evolution and Diversity of a Fungal Self/Nonspecific Recognition Locus. *PLoS ONE*, 5(11). <https://doi.org/10.1371/journal.pone.0014055>
- Hamill, S. J., Steward, A., & Clarke, J. (2000). The Folding of an Immunoglobulin-like Greek Key Protein is Defined by a Common-core Nucleus and Regions Constrained by Topology. *Journal of Molecular Biology*, 297, 165–178. <https://doi.org/10.1006/jmbi.2000.3517>
- Harpaz, Y., & 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. *Journal of Molecular Biology*, 238, 528–539.
- Heinig, M., & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, 32, 500–502. <https://doi.org/10.1093/nar/gkh429>
- Hemmingsen, J. M., Gernert, K. M., Richardson, J. S., & Richardson, D. C. (1994). The tyrosine corner: A feature of most Greek key b-barrel proteins. *Protein Science*, 3, 1927–1937.

- Hirose, S., Benabentos, R., Ho, H.-I., Kuspa, A., & Shaulsky, G. (2011). Self-recognition in social amoebae is mediated by allelic pairs of tiger genes. *Science*, 333(6041), 467–470. <https://doi.org/10.1126/science.1203903>
- Hirose, S., Chen, G., Kuspa, A., & Shaulsky, G. (2017). The polymorphic proteins TgrB1 and TgrC1 function as a ligand–receptor pair in Dictyostelium allorecognition. *The Company of Biologists*, 130, 4002–4012. <https://doi.org/10.1242/jcs.208975>
- Ho, H.-I., Hirose, S., Kuspa, A., & Shaulsky, G. (2013). Kin Recognition Protects Cooperators against Cheaters. *Current Biology*, 23(16), 1590–1595. <https://doi.org/10.1016/j.cub.2013.06.049>
- Ho, H., & Shaulsky, G. (2015). Temporal regulation of kin recognition maintains recognition-cue diversity and suppresses cheating. *Nature Communications*, 6(7144), 1–6. <https://doi.org/10.1038/ncomms8144>
- Hoegh-Guldberg, O., Eakin, C. M., Hodgson, G., Sale, P. F., & Veron, J. E. N. (2018). Consensus Statement on Coral Bleaching Climate Change. *International Coral Reef Society*, 14–17.
- Hoegh-Guldberg, O., Jacob, D., Taylor, M., Guillen Bolanos, T., Bindi, M., Brown, S., Camilloni, I. A., Diedhiou, A., Djalante, R., Ebi, K., Engelbrecht, F., Guiot, J., Hijioka, Y., Mehrotra, S., Hope, C. W., Payne, A. J., Portner, H. O., Seneviratne, S. I., Thomas, A., ... Zhou, G. (2019). The human imperative of stabilizing global climate change at 1.5°C. *Science*, 365(6459). <https://doi.org/10.1126/science.aaw6974>
- Hoft, M. A., Hoving, J. C., & Brown, G. D. (2020). Signaling C-Type Lectin Receptors in Antifungal Immunity. *Current Topics in Microbiology and Immunology*, 429, 63–101. [https://doi.org/https://doi.org/10.1007/82\\_2020\\_224](https://doi.org/https://doi.org/10.1007/82_2020_224)
- Holm, L. (2020). Using Dali for Protein Structure Comparison. In *Structural Bioinformatics: Methods and Protocols* (Vol. 2112, pp. 29–42).
- Hu, E., Bosworth, B., Baxter, J., & Tiersch, T. R. (2014). On-site evaluation of commercial-scale hybrid catfish production using cryopreserved blue catfish sperm. *Aquaculture*, 426–427, 88–95. <https://doi.org/10.1016/j.aquaculture.2014.01.024>
- Hu, E., Childress, W., & Tiersch, T. R. (2017). 3-D printing provides a novel approach for standardization and reproducibility of freezing devices. *Cryobiology*, 76, 34–40. <https://doi.org/10.1016/j.cryobiol.2017.03.010>
- Hu, E., Liao, T. W., & Tiersch, T. R. (2013). A quality assurance initiative for commercial-scale production in high-throughput cryopreservation of blue catfish sperm. *Cryobiology*, 67, 214–224. <https://doi.org/10.1016/j.cryobiol.2013.07.001>
- Hu, E., Liao, T. W., & Tiersch, T. R. (2015). Simulation modeling of high-throughput cryopreservation of aquatic germplasm: a case study of blue catfish sperm processing. *Aquaculture Research*, 46(2), 432–445. <https://doi.org/10.1111/are.12192>

- Hu, E., Yang, H., & Tiersch, T. R. (2011). High-throughput cryopreservation of spermatozoa of blue catfish (*Ictalurus furcatus*): establishment of an approach for commercial-scale processing. *Cryobiology*, 62(1), 74–82. <https://doi.org/10.1016/j.cryobiol.2010.12.006>
- Hua, Z., Fields, A., & Kao, T. (2008). Biochemical Models for S-RNase-Based. *Molecular Plant*, 1(4), 575–585. <https://doi.org/10.1093/mp/ssn032>
- Huene, A. L., Chen, T., & Nicotra, M. L. (2021). New binding specificities evolve via point mutation in an invertebrate allorecognition gene. *IScience*, 24(102811), 1–24. <https://doi.org/10.1016/j.isci.2021.102811>
- Humphreys, T. (1963). Chemical Dissolution and in Vitro Reconstruction of Sponge and Cell Adhesions. *Developmental Biology*, 8, 27–47.
- Humphreys, T. (1970a). Biochemical analysis of sponge cell aggregation. *Symposia of the Zoological Society of London*, 25, 335–352.
- Humphreys, T. (1970b). Species Specific Aggregation of Dissociated Sponge Cells. *Nature*, 228, 685–686.
- Igic, B., Lande, R., & Koh, J. R. (2008). Loss of Self - Incompatibility and Its Evolutionary Consequences. *International Journal of Plant Sciences*, 169(1), 93–104. <https://doi.org/10.1086/523362>
- Iorger, T. R., Clark, A. G., & Kao, T. (1990). Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proceedings of the National Academy of Science*, 87(December), 9732–9735.
- Ivker, F. B. (2014). A Hierarchy of Histo-Incompatibility in *Hydractinia echinata*. *Biological Bulletin*, 143(1), 162–174.
- Iwano, M., & Takayama, S. (2012). Self/non-self discrimination in angiosperm self-incompatibility. *Current Opinion in Plant Biology*, 15(1), 78–83. <https://doi.org/10.1016/j.pbi.2011.09.003>
- James, T. Y. (2015). Why mushrooms have evolved to be so promiscuous: Insights from evolutionary and ecological patterns. *Fungal Biology Reviews*, 29(3–4), 167–178. <https://doi.org/10.1016/j.fbr.2015.10.002>
- Jumblatt, J. E., Schlup, V., & Burger, M. M. (1980). Cell-Cell Recognition: Specific Binding of Microcystin Sponge Aggregation Factor to Homotypic Cells and the Role of Calcium Ions. *Biochemistry*, 19, 1038–1042.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(May), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>

- Kaneko, I., Dementhon, K., Xiang, Q., & Glass, N. L. (2006). Nonallelic Interactions Between het-c and a Polymorphic Locus, pin-c, Are Essential for Nonself Recognition and Programmed Cell Death in *Neurospora crassa*. *Genetics*, 172, 1545–1555. <https://doi.org/10.1534/genetics.105.051490>
- Karadge, U. B., Gosto, M., & Nicotra, M. L. (2015). Allorecognition Proteins in an Invertebrate Exhibit Homophilic Interactions. *Current Biology*, 25(21), 2845–2850. <https://doi.org/10.1016/j.cub.2015.09.030>
- Kasinrerk, W., Tokrasinwit, N., & Phunpae, P. (1999). CD147 monoclonal antibodies induce homotypic cell aggregation of monocytic cell line U937 via LFA-1/ICAM-1 pathway. *Immunology*, 96(2), 184–192. <https://doi.org/10.1046/j.1365-2567.1999.00653.x>
- Katoh, K., Kuma, K., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2), 511–518. <https://doi.org/10.1093/nar/gki198>
- Katoh, K., & Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, 26(15), 1899–1900. <https://doi.org/10.1093/bioinformatics/btq224>
- Katsamba, P., Carroll, K., Ahlsen, G., Bahna, F., Vendome, J., Posy, S., Rajebhosale, M., Price, S., Jessell, T. M., Ben-Shaul, A., Shapiro, L., & Honig, B. H. (2009). Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28), 11594–11599. <https://doi.org/10.1073/pnas.0905349106>
- Kaufman, J. (2018). Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates. *Annual Review of Immunology*, 36, 383–409.
- Kearns, D. B. (2010). A field guide to bacterial swarming motility. *Nature Reviews Microbiology*, 8(September), 634–644. <https://doi.org/10.1038/nrmicro2405>
- Kelley, J., Walter, L., & Trowsdale, J. (2005). Comparative Genomics of Natural Killer Cell Receptor Gene Clusters. *PLoS Genetics*, 1(2), 129–139. <https://doi.org/10.1371/journal.pgen.0010027>
- Kemp, B. P., & Doughty, J. (2007). S cysteine-rich (SCR) binding domain analysis of the Brassica self-incompatibility S-locus receptor kinase. *New Phytologist*, 175, 619–629.
- Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4), 357–360. <https://doi.org/10.1038/nmeth.3317>
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37, 907–915. <https://doi.org/10.1038/s41587-019-0201-4>
- Kolbinger, A., Gao, T., Brock, D., Ammann, R., Kisters, A., Kellermann, J., Hatton, D., Gomer, R. H., & Wetterauer, B. (2005). A Cysteine-Rich Extracellular Protein Containing a PA14

- Domain Mediates Quorum Sensing in *Dictyostelium discoideum*. *Eukaryotic Cell*, 4(6), 991–998. <https://doi.org/10.1128/EC.4.6.991-998.2005>
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27, 722–736. <https://doi.org/10.1101/gr.215087.116>
- Kraus, Y., Flici, H., Hensel, K., Plickert, G., Leitz, T., & Frank, U. (2014). The embryonic development of the cnidarian *Hydractinia echinata*. *Evolution and Development*, 338, 323–338. <https://doi.org/10.1111/ede.12100>
- Krogh, A., Larsson, B., Heijne, G. von, & Sonnhammer, E. L. L. (2001). Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *Journal of Molecular Biology*, 305, 567–580. <https://doi.org/10.1006/jmbi.2000.4315>
- Kubo, Ken-ichi, Entani, T., Takara, A., Wang, N., Fields, A. M., Hua, Z., Toyoda, M., Kawashima, S., Ando, T., Isogai, A., Kao, T., & Takayama, S. (2010). Collaborative Non-Self Recognition System in S-RNase-Based Self-Incompatibility. *Science*, 330, 796–799. <https://doi.org/10.1126/science.1195243>
- Kubo, Ken-inchi, Paape, T., Hatakeyama, M., Entani, T., Takara, A., Kajihara, K., Tsukahara, M., Shimizu-inatsugi, R., Shimizu, K. K., & Takayama, S. (2015). Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in *Petunia*. *Nature Plants*, 1(January). <https://doi.org/10.1038/nplants.2014.5>
- Kubow, K. E., Vukmirovic, R., Zhe, L., Klotzsch, E., Smith, M. L., Gourdon, D., Luna, S., & Vogel, V. (2015). Mechanical forces regulate the interactions of fibronectin and collagen I in extracellular matrix. *Nature Communications*, 6(8026), 1–11. <https://doi.org/10.1038/ncomms9026>
- Kues, U. (2015). From two to many: Multiple mating types in Basidiomycetes. *Fungal Biology Reviews*, 29(3–4), 126–166. <https://doi.org/10.1016/j.fbr.2015.11.001>
- Kues, U., & Casselton, L. A. (1992). Fungal mating type genes - regulators of sexual development. *Mycology Research*, 96(12), 993–1006.
- Kundert, P., & Shaulsky, G. (2019). Cellular allorecognition and its roles in *Dictyostelium* development and social evolution. *International Journal of Developmental Biology*, 393(63), 383–393. <https://doi.org/10.1387/ijdb.190239gs> [www.intjdevbiol.com](http://www.intjdevbiol.com)
- Künzel, T., Heiermann, R., Frank, U., Müller, W., Tilmann, W., Bause, M., Nonn, A., Helling, M., Schwarz, R. S., & Plickert, G. (2010). Migration and differentiation potential of stem cells in the cnidarian *Hydractinia* analysed in eGFP-transgenic animals and chimeras. *Developmental Biology*, 348(1), 120–129. <https://doi.org/10.1016/j.ydbio.2010.08.017>
- Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2),

1–9.

- Kwon, K., & Raper, K. B. (1967). Heterkaryon Formation and Genetic Analyses of Color Mutants in *Aspergillus*. *American Journal of Botany*, 54(1), 49–60.
- Laird, D. J., De Tomaso, A. W., & 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., Tsien, R. Y., & Lin, M. Z. (2012). Improving FRET dynamic range with bright green and red fluorescent proteins. *Nature Methods*, 9(10), 1005–1012. <https://doi.org/10.1038/NMETH.2171>
- Lang, R. P., Riley, K. L., Chandler, J. E., & Tiersch, T. R. (2003). The Use of Dairy Protocols for Sperm Cryopreservation of Blue Catfish *Ictalurus furcatus*. *Journal of the World Aquaculture Society*, 34(1).
- Lanier, L. L. (2008). Up on the tightrope: natural killer cell activation and inhibition. *Nature Immunology*, 9(5), 495–502. <https://doi.org/10.1038/ni1581>
- Lawrence, M. J. (2000). Population Genetics of the Homomorphic Self-incompatibility Polymorphisms in Flowering Plants. *Annals of Botany*, 85, 221–226.
- Leahy, D. J., Hendrickson, W. A., Aukhil, L., & Erickson, H. P. (1992). Structure of a Fibronectin Type III Domain from Tenascin Phased by MAD Analysis of the Selenomethionyl Protein. *Science*, 258(5084), 987–991.
- Lee, H.-S., Huang, S., & Kao, T. (1994). S proteins control rejection of incompatible pollen in *Pteunia inflata*. *Nature*, 367, 560–563.
- Leith, A. (1979). Role of Aggregation Factor and Cell Type in Sponge Cell Adhesion. *Biological Bulletin*, 156, 212–223.
- Lesk, A. M., & Chothia, C. (1982). Evolution of Proteins Formed by  $\beta$ -Sheets II. The Core of the Immunoglobulin Domains. *Journal of Molecular Biology*, 160, 325–342.
- Leslie, J. F. (1993). FUNGAL VEGETATIVE COMPATIBILITY. *Annual Review of Phytopathology*, 31, 127–150.
- Leticia Torres, Hu, E., & Tiersch, T. R. (2016). Cryopreservation in fish: current status and pathways to quality assurance and quality control in repository development. *Reproduction, Fertility and Development*, 28, 1105–1115.
- Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and. *Nucleic Acids Research*, 47, 256–259. <https://doi.org/10.1093/nar/gkz239>
- Letunic, I., Doerks, T., & Bork, P. (2012). SMART 7: recent updates to the protein domain

- annotation resource. *Nucleic Acids Research*, 40, 302–305. <https://doi.org/10.1093/nar/gkr931>
- Li, J., Zhang, Y., Song, Y., Zhang, H., Fan, J., Li, Q., Zhang, D., & Xue, Y. (2017). Electrostatic potentials of the S -locus F-box proteins contribute to the pollen S specificity in self-incompatibility in *Petunia hybrida*. *The Plant Journal*, 89, 45–57. <https://doi.org/10.1111/tpj.13318>
- Litman, G. W., Hawke, N. A., & Yoder, J. A. (2001). Novel immune-type receptor genes. *Immunological Reviews*, 181, 250–259.
- Liu, Y., Torres, L., Tiersch, T. R., & Rouge, B. (2018). Quality evaluation of sperm from livebearing fishes: standardized assessment of sperm bundles (spermatozeugmata) from *Xenotoca eiseni* (Goodeidae). *Theriogenology*, 107, 50–56. <https://doi.org/10.1016/j.theriogenology.2017.10.037>
- Liu, Y., Yang, H., Torres, L., & Tiersch, T. R. (2018). Activation of free sperm and dissociation of sperm bundles (spermatozeugmata) of an endangered viviparous Fish, *Xenotoca eiseni*. *Comp Biochem Physiol A Mol Integr Physiol*, 218, 38–45. <https://doi.org/10.1016/j.cbpa.2018.01.006>
- Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology (Clifton, N.J.)*, 1079, 155–170. [https://doi.org/10.1007/978-1-62703-646-7\\_10](https://doi.org/10.1007/978-1-62703-646-7_10)
- Lyons, N. A., Kraigher, B., Stefanic, P., Mandic-Mulec, I., & Kolter, R. (2016). A Combinatorial Kin Discrimination System in *Bacillus subtilis*. *Current Biology*, 26(6), 733–742. <https://doi.org/10.1016/j.cub.2016.01.032>
- Mali, B., Millane, R. C., Plickert, G., Frohme, M., & Frank, U. (2011). A polymorphic, thrombospondin domain-containing lectin is an oocyte marker in *Hydractinia*: implications for germ cell specification and sex determination. *International Journal of Developmental Biology*, 55, 103–108. <https://doi.org/10.1387/ijdb.103063bm>
- Matton, D. P., Luu, D. T., Xike, Q., Laublin, G., Brien, M. O., Maes, O., Morse, D., & Cappadocia, M. (1999). Production of an S RNase with Dual Specificity Suggests a Novel Hypothesis for the Generation of New S Alleles. *The Plant Cell*, 11(November), 2087–2097.
- McClay, D. R. (1971). An Autoradiographic Analysis of the Species Specificity during Sponge Cell Reaggregation. *Biological Bulletin*, 141(2), 319–330.
- McClure, B. (2009). Darwin's foundation for investigating self-incompatibility and the progress toward a physiological model for S-RNase-based SI. *Journal of Experimental Biology*, 60(4), 1069–1081. <https://doi.org/10.1093/jxb/erp024>
- McKittrick, T. R., & De Tomaso, A. W. (2010). Molecular mechanisms of allorecognition in a basal chordate. *Seminars in Immunology*, 22, 34–38. <https://doi.org/10.1016/j.smim.2009.12.001>

- McKittrick, T. R., Muscat, C. C., Pierce, J. D., Bhattacharya, D., & De Tomaso, A. W. (2011). Allorecognition in a Basal Chordate Consists of Independent Activating and Inhibitory Pathways. *Immunity*, 34(4), 616–626. <https://doi.org/10.1016/j.immuni.2011.01.019>
- Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., & Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, 7(339), 1–10. <https://doi.org/10.1186/1471-2105-7-339>
- Meng, X., Sun, P., & Kao, T. (2011). S-RNase-based self-incompatibility in *Petunia inflata*. *Annals of Botany*, 108, 637–646. <https://doi.org/10.1093/aob/mcq253>
- Metzenberg, R. L., & Glass, N. L. (1990). Mating Type and Mating Strategies in *Neurospora*. *BioEssays*, 12(2), 53–59.
- Milkman, R. (1967). Genetic and Developmental Studies on *Botryllus schlosseri*. *Biological Bulletin*, 132(2), 229–243.
- Mirdita, M., Ovchinnikov, S., & Steinegger, M. (2021). ColabFold - Making protein folding accessible to all. *Preprint*, 4–9. <https://doi.org/https://doi.org/10.1101/2021.08.15.456425>
- Mócsai, A., Ruland, J., & Tybulewicz, V. L. J. (2010). The SYK tyrosine kinase: a crucial player in diverse biological functions. *Nature Reviews Immunology*, 10, 387–402. <https://doi.org/10.1038/nri2765>
- Mokady, O., & Buss, L. W. (1996). Transmission genetics of allorecognition in *Hydractinia symbiolongicarpus* (Cnidaria: Hydrozoa). *Genetics*, 143(2), 823–827.
- Mueller, W A. (1964). Experimental investigations on colony development, polyp differentiation and sexual chimeras in *Hydractinia echinata* (Foreign title: Experimentelle Untersuchungen über Stockentwicklung, Polypendifferenzierung und Sexualchimären bei *Hydractinia echinata*). *Wilhelm Roux' Arch Für Entwick- Lungsmechanik*, 155, 181–268.
- Mueller, Werner A, & Rinkevich, B. (2020). Cell Communication-mediated Nonself-Recognition and -Intolerance in Representative Species of the Animal Kingdom. *Journal of Molecular Evolution*, 88(6), 482–500. <https://doi.org/10.1007/s00239-020-09955-z>
- Muirhead, C. A., Glass, N. L., & Slatkin, M. (2002). Multilocus Self-Recognition Systems in Fungi as a Cause of Trans-Species Polymorphism. *Genetics*, 641(June), 633–641.
- Mukai, H., & Watanabe, H. (1975). Distribution of Fusion Incompatibility Types in Natural Populations of the Compound Ascidian, *Botryllus primigenus*. *Proceedings of the Japan Academy*, 51(289), 44–47.
- Muller, W. A. (1973). Induction of Metamorphosis by Bacteria and Ions in the Planulae of *Hydractinia Echinata*; An Approach to the Mode of Action. *Publications of the Seto Marine Biological Laboratory*.
- Muñoz-sanz, J. V., Zuriaga, E., Cruz-garcía, F., McClure, B., & Romero, C. (2020). Self-



- (In)compatibility Systems: Target Traits for Crop-Production, Plant Breeding, and Biotechnology. *Frontiers in Plant Science*, 11(195). <https://doi.org/10.3389/fpls.2020.00195>
- Murfett, J., Atherton, T. L., Mou, B., Gasser, C. S., & McClure, B. A. (1994). S-RNase expressed in transgenic *Nicotiana* causes S-allele-specific pollen rejection. *Nature*, 367, 563–566.
- Murphy, K., Travers, P., Walport, M., & Janeway, C. (2008). *Janeway's Immunobiology* (7th ed.). Garland Science.
- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Pond, S. L. K. (2012). Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genetics*, 8(7), 1–10. <https://doi.org/10.1371/journal.pgen.1002764>
- Mylyk, O. M. (1975). Heterokaryon Incompatibility Genes in *Neurospora Crassa* Detected Using Duplication-Producing Chromosome Rearrangements. *Genetics*, 80(May), 104–124.
- Naithani, S., Chookajorn, T., Ripoll, D. R., & Nasrallah, J. B. (2007). Structural modules for receptor dimerization in the S-locus receptor kinase extracellular domain. *Proceedings of the National Academy of Science*, 104(29), 12211–12216.
- Nasrallah, J. B. (2019). Self-incompatibility in the Brassicaceae: Regulation and mechanism of self-recognition. In *Plant Development and Evolution* (1st ed., Vol. 131, pp. 435–452). Elsevier Inc. <https://doi.org/10.1016/bs.ctdb.2018.10.002>
- Newbigin, E., Paape, T., & Kohn, J. R. (2008). RNase-Based Self-Incompatibility: Puzzled by Pollen S. *The Plant Cell*, 20(September), 2286–2292. <https://doi.org/10.1105/tpc.108.060327>
- Newmeyer, D., Howe, H. B. J., & Galeazzi, D. R. (1973). A SEARCH FOR COMPLEXITY AT THE MATING-TYPE LOCUS OF *NEUROSPORA CRASSA*. *Canadian Journal of Genetics and Cytology*, 15(3), 577–585.
- Nicotra, M. L. (2019). Invertebrate allorecognition. *Current Biology*, 29(11), R463–R467. <https://doi.org/10.1016/j.cub.2019.03.039>
- Nicotra, M. L., & Buss, L. W. (2005). A test for larval kin aggregations. *Biological Bulletin*, 208(3), 157–158.
- Nicotra, M. L., Powell, A. E., Rosengarten, R. D., Moreno, M., Grimwood, J., Lakkis, F. G., Dellaporta, S. L., & Buss, L. W. (2009). A Hypervariable Invertebrate Allodeterminant. *Current Biology*, 19(7), 583–589. <https://doi.org/10.1016/j.cub.2009.02.040>
- Noe, L., & Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research*, 33, 540–543. <https://doi.org/10.1093/nar/gki478>
- Nydam, M. L., & De Tomaso, A. W. (2012). The fester locus in *Botryllus schlosseri* experiences selection. *BMC Evolutionary Biology*, 12(249), 1–16.
- Nydam, M. L., Netuschil, N., Sanders, E., Langenbacher, A., Lewis, D. D., Taketa, D. A.,

- Marimuthu, A., Gracey, A. Y., & De Tomaso, A. W. (2013). The Candidate Histocompatibility Locus of a Basal Chordate Encodes Two Highly Polymorphic Proteins. *PLoS ONE*, 8(6), 1–13. <https://doi.org/10.1371/journal.pone.0065980>
- Nydam, M. L., Stephenson, E. E., Waldman, C. E., & De Tomaso, A. W. (2017). Balancing selection on allorecognition genes in the colonial ascidian *Botryllus schlosseri*. *Developmental and Comparative Immunology*, 69, 60–74. <https://doi.org/10.1016/j.dci.2016.12.006>
- Nyholm, S. V., Passegue, E., Ludington, W. B., Voskoboynik, A., Mitchel, K., Weissman, I. L., & De Tomaso, A. W. (2006). *fester*, a Candidate Allorecognition Receptor from a Primitive Chordate. *Immunity*, 25, 163–173. <https://doi.org/10.1016/j.immuni.2006.04.011>
- Ohki, S., Morita, M., Kitanobo, S., Kowalska, A. A., & Kajetan, R. K. (2014). Cryopreservation of *Acropora digitifera* sperm with use of sucrose and methanol based solution. *Cryobiology*, 69(1), 134–139. <https://doi.org/10.1016/j.cryobiol.2014.06.005>
- Ohm, R. A., Jong, J. F. De, Lugones, L. G., Aerts, A., Kothe, E., Stajich, J. E., Vries, R. P. De, Record, E., Levasseur, A., Baker, S. E., Bartholomew, K. A., Coutinho, P. M., Erdmann, S., Fowler, T. J., Gathman, A. C., Lombard, V., Henrissat, B., Knabe, N., Kües, U., ... Wösten, H. A. B. (2010). Genome sequence of the model mushroom *Schizophyllum commune*. *Nature Biotechnology*, 28(9), 957–965. <https://doi.org/10.1038/nbt.1643>
- Oka, H., & Watanabe, H. (1960). Problems of Colony-Specificity in Compound Ascidians. *Bulletin Marine Biology Station of Asamushi*, 10(2), 153–155.
- Ostrowski, E. A., Katoh, M., Shaulsky, G., Queller, D. C., & Strassmann, J. E. (2008). Kin Discrimination Increases with Genetic Distance in a Social Amoeba. *PloS Biology*, 6(11), 2376–2382. <https://doi.org/10.1371/journal.pbio.0060287>
- Palumbi, S. R., Anthony, K. R. N., Baker, A. C., Baskett, M. L., Bhattacharya, D., Bourne, D. G., Knowlton, N., Logan, C. A., Naish, K. A., Richmond, R. H., Smith, T. B., & von Stackelberg, K. (2019). A Research Review of Interventions to Increase the Persistence and Resilience of Coral Reefs. In *A Research Review of Interventions to Increase the Persistence and Resilience of Coral Reefs*. The National Academies Press. <https://doi.org/10.17226/25279>
- Paniagua-chavez, C. G., Buchanan, J. T., Supan, J. E., & Tiersch, T. R. (2000). Cryopreservation of Sperm and Larvae of the Eastern Oyster. *Cryopreservation in Aquatic Species*, 230–239.
- Paoletti, M., Saupe, S. J., & Clave, C. (2007). Genesis of a Fungal Non-Self Recognition Repertoire. *PLoS ONE*, 3. <https://doi.org/10.1371/journal.pone.0000283>
- Partridge, J. D., Ariel, G., Schvartz, O., Harshey, R. M., & Be'er, A. (2018). The 3D architecture of a bacterial swarm has implications for antibiotic tolerance. *Scientific Reports*, 8(15823), 1–11. <https://doi.org/10.1038/s41598-018-34192-2>
- Partridge, J. D., & Harshey, R. M. (2013). Swarming: Flexible Roaming Plans. *Journal of Bacteriology*, 195(5), 909–918. <https://doi.org/10.1128/JB.02063-12>

- Pathak, D. T., Wei, X., Bucuvalas, A., Haft, D. H., Gerloff, D. L., & Wall, D. (2012). Cell Contact–Dependent Outer Membrane Exchange in Myxobacteria: Genetic Determinants and Mechanism. *PLoS Genetics*, 8(4), 1–12. <https://doi.org/10.1371/journal.pgen.1002626>
- Pathak, D. T., Wei, X., Dey, A., & Wall, D. (2013). Molecular Recognition by a Polymorphic Cell Surface Receptor Governs Cooperative Behaviors in Bacteria. *PLoS Genetics*, 9(11), 1–12. <https://doi.org/10.1371/journal.pgen.1003891>
- Perkins, D. D. (1988). Main features of vegetative incompatibility in *Neurospora*. *Fungal Genetics Reports*, 35, 4–7.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera — A Visualization System for Exploratory Research and Analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612. <https://doi.org/10.1002/jcc.20084>
- Phillee, M. L., & Staben, C. (1994). Functional Analyses of the *Neurospora crassa* MT a-1 Mating Type Polypeptide. *Genetics*, 137, 713–722.
- Plickert, G., Kroiher, M., & Munck, A. (1988). Cell proliferation and early differentiation during embryonic development and metamorphosis of *Hydractinia echinata*. *Development*, 103, 795–803.
- Plickert, Günter, Frank, U., & Müller, W. A. (2012). *Hydractinia*, a pioneering model for stem cell biology and reprogramming somatic cells to pluripotency. *International Journal of Developmental Biology*, 56, 519–534. <https://doi.org/10.1387/ijdb.123502gp>
- Podgornaia, A. I., & Laub, M. T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222), 673–678.
- Pond, S. L. K., & Frost, S. D. W. (2005). Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution*, 22(5), 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., Wisotsky, S., Spielman, S. J., Frost, S. D. W., & Muse, S. V. (2019). HyPhy 2.5 — A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, 37(1), 295–299. <https://doi.org/10.1093/molbev/msz197>
- Povolotskaya, I. S., & Kondrashov, F. A. (2010). Sequence space and the ongoing expansion of the protein universe. *Nature*, 465(7300), 922–926. <https://doi.org/10.1038/nature09105>
- Powell, A. E., Moreno, M., Gloria-soria, A., Lakkis, F. G., Dellaporta, S. L., & Buss, L. W. (2011). Genetic Background and Allorecognition Phenotype in *Hydractinia symbiolongicarpus*. *G3*, 1(November), 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., & Buss, L. W.

- (2007). Differential effect of allorecognition loci on phenotype in *Hydractinia symbiolongicarpus* (Cnidaria: Hydrozoa). *Genetics*, 177(4), 2101–2107. <https://doi.org/10.1534/genetics.107.075689>
- R. Lange, G. P. W. A. M. (1989). Histoincompatibility in a low invertebrate, *Hydractinia echinata*: Analysis of the mechanism of rejection. *Journal of Experimental Zoology*, 249(3), 284–292.
- Raper, J. R., Krongelb, G. S., & Baxter, M. G. (1958). THE NUMBER AND DISTRIBUTION OF INCOMPATIBILITY FACTORS IN SCHIZOPHYLLUM. *The American Naturalist*, XCII(865), 221–232.
- Rea, A. C., & Nasrallah, J. B. (2008). Self-incompatibility systems: barriers to self-fertilization in flowering plants. *International Journal of Developmental Biology*, 52, 627–636. <https://doi.org/10.1387/ijdb.072537ar>
- Richardson, J. S. (1981). The Anatomy and Taxonomy of Protein Structure. In *Advances in Protein Chemistry* (Vol. 34, pp. 167–339).
- Richman, A. (2000). Evolution of balanced genetic polymorphism. *Molecular Ecology*, 9, 1953–1963.
- Richman, A. D., & Kohn, J. R. (2000). Evolutionary genetics of self-incompatibility in the Solanaceae. *Plant Molecular Biology*, 42, 169–179.
- Rigden, D. J., Mello, L. V., & Galperin, M. Y. (2004). The PA14 domain, a conserved all-beta domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends in Biochemical Sciences*, 29(7), 335–339.
- Rinkevich, B., Shapira, M., Weissman, I. L., & Saito, Y. (1992). Allogeneic Responses between Three Remote Populations of the Cosmopolitan Ascidian *Botryllus schlosseri*. *Zoological Science*, 9, 989–994.
- Rosa, S. F. P., Powell, A. E., Rosengarten, R. D., Nicotra, M. L., Moreno, M. A., Grimwood, J., Lakkis, F. G., Dellaporta, S. L., & Buss, L. W. (2010). *Hydractinia* allodeterminant alr1 resides in an immunoglobulin superfamily-like gene complex. *Current Biology*, 20(12), 1122–1127. <https://doi.org/10.1016/j.cub.2010.04.050>
- Rosengarten, R. D., Moreno, M. A., Lakkis, F. G., Buss, L. W., & Dellaporta, S. L. (2011). Genetic diversity of the allodeterminant alr2 in *hydractinia symbiolongicarpus*. *Molecular Biology and Evolution*, 28(2), 933–947. <https://doi.org/10.1093/molbev/msq282>
- Rosengarten, R. D., & Nicotra, M. L. (2011). Model systems of invertebrate allorecognition. *Current Biology*, 21(2), R82–R92. <https://doi.org/10.1016/j.cub.2010.11.061>
- Rubinstein, R., Thu, C. A., Goodman, K. M., Wolcott, H. N., Bahna, F., Mannepalli, S., Ahlsen, G., Chevee, M., Halim, A., Clausen, H., Maniatis, T., Shapiro, L., & Honig, B. (2015). Molecular Logic of Neuronal Self-Recognition through Protocadherin Domain Interactions. *Cell*, 163(3), 629–642. <https://doi.org/10.1016/j.cell.2015.09.026>

- Saak, C. C., & Gibbs, K. A. (2016). The Self-Identity Protein IdsD Is Communicated between Cells in Swarming *Proteus mirabilis* Colonies. *Journal of Bacteriology*, 198(24), 3278–3286. <https://doi.org/10.1128/JB.00402-16>
- Sanders, S. M., Ma, Z., Hughes, J. M., Riscoe, B. M., Gibson, G. A., Watson, A. M., Flici, H., Frank, U., Schnitzler, C. E., Baxeavanis, A. D., & Nicotra, M. L. (2018). CRISPR/Cas9-mediated gene knockin in the hydroid *Hydractinia symbiolongicarpus*. *BMC Genomics*, 19(1), 1–17. <https://doi.org/10.1186/s12864-018-5032-z>
- Sanders, S. M., Shcheglovitova, M., & Cartwright, P. (2014). Differential gene expression between functionally specialized polyps of the colonial hydrozoan *Hydractinia symbiolongicarpus* (Phylum Cnidaria). *BMC Genomics*, 15(1), 406. <https://doi.org/10.1186/1471-2164-15-406>
- Saupe, S. J. (2000). Molecular Genetics of Heterokaryon Incompatibility in Filamentous Ascomycetes. *Microbiology and Molecular Biology Reviews*, 64(3), 489–502.
- Saupe, S., Turcq, B., & Begueret, J. (1995). Sequence diversity and unusual variability at the *het-c* locus involved in vegetative incompatibility in the fungus *Podospira anserina*. *Current Genetics*, 27, 466–471.
- Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7), 671–675. <https://doi.org/10.1038/nmeth.2089>
- Schnitzler, C. E., Baxeavanis, A. D., Nicotra, M. L., & Frank, U. (2017). *Hydractinia Genome Project Portal*. <https://research.nhgri.nih.gov/hydractinia/>
- Schopfer, C. R., Nasrallah, M. E., & Nasrallah, J. B. (1999). The Male Determinant of Self-Incompatibility in Brassica. *Science*, 286(November), 1697–1701.
- Schreiner, D., & Weiner, J. a. (2010). Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33), 14893–14898. <https://doi.org/10.1073/pnas.1004526107>
- Scofield, V L, Schlumpberger, J. M., West, L. A., & Weissman, I. L. (1982). Protochordate allorecognition is controlled by a MHC-like gene system. *Nature*, 295, 499–502.
- Scofield, Virginia L, Schlumpberger, J. M., & Weissman, I. L. (1982). Colony Specificity in the Colonial Tunicate *Botryllus* and the Origins of Vertebrate Immunity. *American Zoology*, 22, 783–794.
- Senior, B. W. (1977). The Dienes Phenomenon: Identification of the Determinants of Compatibility. *Journal of General Microbiology*, 102, 235–244.
- Shamkhalichenar, H., Choi, J., & Tiersch, T. R. (2019). Three-dimensional printing can provide customizable probes for sensing and monitoring in cryobiology applications. *Cryobiology*, 88, 64–69. <https://doi.org/10.1016/j.cryobiol.2019.03.010>

- Shiba, H., Iwano, M., Entani, T., Ishimoto, K., Shimosato, H., Che, F.-S., Satta, Y., Ito, A., Takada, Y., Watanabe, M., Isogai, A., & Takayama, S. (2002). The Dominance of Alleles Controlling Self-Incompatibility in Brassica Pollen Is Regulated at the RNA Level. *The Plant Cell*, 14(February), 491–504. <https://doi.org/10.1105/tpc.010378>
- Shiu, P. K. T., & Glass, N. L. (1999). Molecular Characterization of tol, a Mediator of Mating-Type-Associated Vegetative Incompatibility in *Neurospora crassa*. *Genetics*, 151, 545–555.
- Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). Genome analysis BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Smith, M. L., Gourdon, D., Little, W. C., Kubow, K. E., Eguiluz, R. A., Luna-Morris, S., & Vogel, V. (2007). Force-Induced Unfolding of Fibronectin in the Extracellular Matrix of Living Cells. *PLoS Biology*, 5(10), 2243–2254. <https://doi.org/10.1371/journal.pbio.0050268>
- Specht, C. A. (1995). Isolation of the B-alpha and B-beta mating-type loci of *Schizophyllum commune*. *Current Genetics*, 28, 374–379.
- Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M., & Pond, S. L. K. (2019). Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. In *Evolutionary Genomics. Methods in Molecular Biology*. (Vol. 1910, pp. 427–468). <https://doi.org/10.1007/978-1-4939-9074-0>
- Spindler, K., & Werner, A. M. (1972). Induction of Metamorphosis by Bacteria and by a Lithium-pulse in the Larvae of *Hydractinia echinata* (Hydrozoa). *Wilhelm Roux' Archiv*, 280, 271–280.
- Staben, C., & Yanofsky, C. (1990). *Neurospora crassa* a mating-type region. *Proceedings of the National Academy of Science*, 87(July), 4917–4921.
- Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32, 309–312. <https://doi.org/10.1093/nar/gkh379>
- Steele, R. E., Stover, N. A., & Sakaguchi, M. (1999). Appearance and disappearance of Syk family protein-tyrosine kinase genes during metazoan evolution. *Gene*, 239, 91–97.
- Stefanic, P., Kraigher, B., Anthony, N., Kolter, R., & Mandic-Mulec, I. (2015). Kin discrimination between sympatric *Bacillus subtilis* isolates. *Proceedings of the National Academy of Sciences*, 112(45), 14042–14047. <https://doi.org/10.1073/pnas.1512671112>
- Stein, J. C., Howlett, B., Boyes, D. C., Nasrallah, M. E., & Nasrallah, J. B. (1991). Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleracea*. *Proceedings of the National Academy of Sciences USA*, 88(October), 8816–8820.

- Stone, S. L., Amoldo, M., & Goring, D. R. (1999). A Breakdown of Brassica Self-Incompatibility in ARC1 Antisense Transgenic Plants. *Science*, 286, 1729–1731.
- Stone, S. L., Anderson, E. M., Mullen, R. T., & Goring, D. R. (2003). ARC1 Is an E3 Ubiquitin Ligase and Promotes the Ubiquitination of Proteins during the Rejection of Self-Incompatible Brassica Pollen. *The Plant Cell*, 15, 885–898. <https://doi.org/10.1105/tpc.009845>
- Stoner, D. S., Rinkevich, B., & Weissman, I. L. (1999). Heritable germ and somatic cell lineage competitions in chimeric colonial protochordates. *Proc*, 96, 9148–9153.
- Stoner, D. S., & 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*, 93, 15254–15259.
- Sun, J., Wang, L., Yang, W., Wang, L., Fu, Q., & Song, L. (2020). IgIT-Mediated Signaling Inhibits the Antimicrobial Immune Response in Oyster Hemocytes. *The Journal of Immunology*, 205, 1–12. <https://doi.org/10.4049/jimmunol.2000294>
- Sun, M., Wang, Y., Zhang, Q., Xia, Y., Ge, W., & Guo, D. (2017). Prediction of reversible disulfide based on features from local structural signatures. *BMC Genomics*, 18(279), 1–10. <https://doi.org/10.1186/s12864-017-3668-8>
- Sun, P., Li, S., Lu, D., Williams, J. S., & Kao, T. (2015). Pollen S – locus F – box proteins of Petunia involved in S – RNase-based self-incompatibility are themselves subject to ubiquitin-mediated degradation. *The Plant Journal*, 83, 213–223. <https://doi.org/10.1111/tpj.12880>
- Suzuki, G., Kai, N., Hirose, T., Fukui, K., Nishio, T., Takayama, S., Isogai, A., Watanabe, M., & Hinata, K. (1999). Genomic Organization of the S Locus: Identification and Characterization of Genes in SLG/SRK Region of S9 Haplotype of Brassica campestris (syn . rapa). *Genetics*, 153, 391–400.
- Takayama, S., & Isogai, A. (2005). Self-Incompatibility in Plants. *Annual Review of Plant Biology*, 56(1), 467–489. <https://doi.org/10.1146/annurev.arplant.56.032604.144249>
- Takayama, S., Shiba, H., Iwano, M., Shimosato, H., Che, F.-S., Kai, N., Watanabe, M., Suzuki, G., Hinata, K., & Isogai, A. (2000). The pollen determinant of self-incompatibility in Brassica campestris. *Proceedings of the National Academy of Science*, 97(4), 1920–1925.
- Taketa, D. A., & De Tomaso, A. W. (2015). Botryllus schlosseri Allorecognition: Tackling the Enigma. *Developmental and Comparative Immunology*, 48(1), 254–265. <https://doi.org/10.1016/j.dci.2014.03.014>
- Thomas, S. G., & Franklin-Tong, V. E. (2004). Self-incompatibility triggers programmed cell death in Papaver pollen. *Nature*, 429(May), 305–309. <https://doi.org/10.1105/tpc.016154>
- Thu, C. A., Chen, W. V., Rubinstein, R., Chevee, M., Wolcott, H. N., Felsovalyi, K. O., Tapia, J. C., Shapiro, L., Honig, B., & Maniatis, T. (2014). Single-Cell Identity Generated by Combinatorial Homophilic Interactions between alpha, beta, and gamma Protocadherins. *Cell*,

158(5), 1045–1059. <https://doi.org/10.1016/j.cell.2014.07.012>

- Tiersch, T. R. (2011a). Process pathways for cryopreservation research, application and commercialization. In T. R. Tiersch & C. C. Green (Eds.), *Cryopreservation in Aquatic Species* (2nd ed., pp. 646–671). World Aquaculture Society.
- Tiersch, T. R. (2011b). *Process Pathways for Cryopreservation Research, Application and Commercialization*. 646–671.
- Tiersch, T. R., & Monroe, W. T. (2016). Three-dimensional printing with polylactic acid (PLA) thermoplastic offers new opportunities for cryobiology Terrence. *Cryobiology*, 73(3), 396–398. <https://doi.org/10.1016/j.cryobiol.2016.10.005>
- Tipping, M. J., & Gibbs, K. A. (2019). Peer pressure from a *Proteus mirabilis* self-recognition system controls participation in cooperative swarm motility. *PLoS Pathogens*, 1–24.
- Torres, L., & Tiersch, T. R. (2018). Addressing Reproducibility in Cryopreservation , and Considerations Necessary for Commercialization and Community Development in Support of Genetic Resources of Aquatic Species. *Journal of the World Aquaculture Society*, 49(4), 644–663. <https://doi.org/10.1111/jwas.12541>
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), 511–515. <https://doi.org/10.1038/nbt.1621>
- Trifinopoulos, J., Nguyen, L., Haeseler, A. Von, & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44, 232–235. <https://doi.org/10.1093/nar/gkw256>
- Troselj, V., Cao, P., & Wall, D. (2018). Cell-cell recognition and social networking in bacteria. *Environmental Microbiology*, 20(3), 923–933. <https://doi.org/10.1111/1462-2920.14005>
- Trowsdale, J., Jones, D. C., Barrow, A. D., & Traherne, J. A. (2015). Surveillance of cell and tissue perturbation by receptors in the LRC. *Immunological Reviews*, 267, 117–136.
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., ... Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596, 590–596. <https://doi.org/10.1038/s41586-021-03828-1>
- Ushijima, K., Sassa, H., Dandekar, A. M., Gradziel, T. M., Tao, R., & Hirano, H. (2003). Structural and Transcriptional Analysis of the Self-Incompatibility Locus of Almond: Identification of a Pollen-Expressed F-Box Gene with Haplotype-Specific Polymorphism. *The Plant Cell*, 15(March), 771–781. <https://doi.org/10.1105/tpc.009290>
- Uyenoyama, M. K., Zhang, Y., & Newbigin, E. (2001). On the Origin of Self-Incompatibility



- Haplotypes : Transition Through Self-Compatible Intermediates. *Genetics*, 157, 1805–1817.
- Van der Nest, M. A., Olson, Å., Lind, M., Vélèz, H., Dalman, K., Durling, M. B., Karlsson, M., & Stenlid, J. (2014). Distribution and evolution of het gene homologs in the basidiomycota. *Fungal Genetics and Biology*, 64, 45–57. <https://doi.org/10.1016/j.fgb.2013.12.007>
- van Sorge, N. M., Bonsor, D. A., Deng, L., Lindahl, E., Schmitt, V., Lyndin, M., Schmidt, A., Nilsson, O. R., Brizuela, J., Boero, E., Sundberg, E. J., van Strijp, J. A. G., Doran, K. S., & Singer, B. B. (2021). Bacterial protein domains with a novel Ig-like fold target human CEACAM receptors. *The EMBO Journal*, 40(e106103). <https://doi.org/10.15252/emboj.2020106103>
- Vassallo, C., Pathak, D. T., Cao, P., Zuckerman, D. M., Hoiczky, E., & Wall, D. (2015). Cell rejuvenation and social behaviors promoted by LPS exchange in myxobacteria. *Proceedings of the National Academy of Science*. <https://doi.org/10.1073/pnas.1503553112>
- Vellani, T. S., Griffiths, A. J. F., & Glass, N. L. (1994). New mutations that suppress mating-type vegetative incompatibility in *Neurospora crassa*. *Genome*, 37(2), 249–255.
- Vieira, J., Rocha, S., Vázquez, N., & López-fernández, H. (2019). Predicting Specificities Under the Non-self Gametophytic Self-Incompatibility Recognition Model. *Frontiers in Plant Science*, 10(July), 1–15. <https://doi.org/10.3389/fpls.2019.00879>
- Viyakarn, V., Chavanich, S., Chong, G., Tsai, S., & Lin, C. (2018). Cryopreservation of sperm from the coral *Acropora humilis*. *Cryobiology*, 80, 130–138. <https://doi.org/10.1016/j.cryobiol.2017.10.007>
- Vos, M., & Velicer, G. J. (2009). Report Social Conflict in Centimeter- and Global-Scale Populations of the Bacterium *Myxococcus xanthus*. *Current Biology*, 19(20), 1763–1767. <https://doi.org/10.1016/j.cub.2009.08.061>
- 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., Dimov, I. K., Keasar, C., Fan, H. C., Mantalas, G. L., Sinha, R., Penland, L., Quake, S. R., & Weissman, I. L. (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., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, 9(11), 1–14. <https://doi.org/10.1371/journal.pone.0112963>
- Wang, J. H. (2013). The sequence signature of an Ig-fold. *Protein and Cell*, 4(8), 569–572. <https://doi.org/10.1007/s13238-013-3903-2>
- Wang, J., Hou, L., Awrey, D., Loomis, W. F., Firtel, R. A., & Siu, C. (2000). The Membrane Glycoprotein gp150 Is Encoded by the lagC Gene and Mediates Cell–Cell Adhesion by Heterophilic Binding during Dictyostelium Development. *Developmental Biology*, 227, 734–745. <https://doi.org/10.1006/dbio.2000.9881>

- Wang, L., Lin, Z., Triviño, M., Nowack, M. K., Franklin-Tong, V. E., & Bosch, M. (2019). Self-incompatibility in Papaver Pollen: Programmed Cell Death in an Acidic Environment. *Journal of Experimental Botany*, 70(7), 2113–2123. <https://doi.org/10.1093/jxb/ery406>
- Watanabe, M., Ito, A., Takada, Y., Ninomiya, C., Kakizaki, T., Takahata, Y., Hatakeyama, K., Hinata, K., Suzuki, G., Takasaki, T., Satta, Y., Shiba, H., Takayama, S., & Isogai, A. (2000). Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of *Brassica campestris* (syn . *rapa*) L. *FEBS Letters*, 473(2), 139–144.
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), 1189–1191. <https://doi.org/10.1093/bioinformatics/btp033>
- Weinreich, D. M., Delaney, N. F., DePristo, M. A., & 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>
- Weis, V. M. (2019). Cell Biology of Coral Symbiosis: Foundational Study Can Inform Solutions to the Coral Reef Crisis. *Integrative and Comparative Biology*, 1–11. <https://doi.org/10.1093/icb/icz067>
- Weis, V. M., Keene, D. R., & Buss, L. E. O. W. (1985). Biology of hydractiniid hydroids. 4. ultrastructure the planula of hydractinia echinata. *Biological Bulletin*, 168, 403–418.
- Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., Russ, C., Nusbaum, C., Lander, E. S., Maccallum, I., & Jaffe, D. B. (2014). Comprehensive variation discovery in single human genomes. *Nature Publishing Group*, 46(12), 1350–1355. <https://doi.org/10.1038/ng.3121>
- Weissberger, E. J. (1995). Association of the Hermit Crab *Pagurus longicarpus* Say, 1817, with Symbiotic Hydroids: Consequences of Predation by Lobsters. *Crustaceana*, 68(6), 739–750.
- Wenren, L. M., Sullivan, N. L., Cardarelli, L., Septer, A. N., & Gibbs, K. A. (2013). Two Independent Pathways for Self-Recognition in *Proteus mirabilis* Are Linked by Type VI-Dependent Export. *MBio*, 4(4), 1–10. <https://doi.org/10.1128/mBio.00374-13>
- Wheeler, D., & Newbiggin, E. (2007). Implications for Understanding the Pollen Factor of the S Locus. *Genetics*, 177, 2171–2180. <https://doi.org/10.1534/genetics.107.076885>
- Wheeler, M. J., Graaf, B. H. J. De, Hadjiosif, N., Perry, R. M., Poulter, N. S., Osman, K., Vatovec, S., Harper, A., Franklin, F. C. H., & Franklin-Tong, V. E. (2009). Identification of the pollen self-incompatibility determinant in *Papaver rhoeas*. *Nature*, 459(7249), 992–995. <https://doi.org/10.1038/nature08027>
- Wheeler, M. J., Vatovec, S., & Franklin-tong, V. E. (2010). The pollen S-determinant in *Papaver*: comparisons with known plant receptors and protein ligand partners. *Journal of Experimental Botany*, 61(7), 2015–2025. <https://doi.org/10.1093/jxb/erp383>

- Williams, J. S., Der, J. P., DePamphilis, C. W., & Kao, T. (2014). Transcriptome Analysis Reveals the Same 17 S-Locus F-Box Genes in Two Haplotypes of the Self-Incompatibility Locus of *Petunia inflata*. *The Plant Cell*, 26(July), 2873–2888. <https://doi.org/10.1105/tpc.114.126920>
- Williams, J. S., Wu, L., Li, S., Sun, P., & Kao, T. (2015). Insight into S-RNase-based self-incompatibility in *Petunia*: recent findings and future directions. *Frontiers in Plant Science*, 6(February), 1–6. <https://doi.org/10.3389/fpls.2015.00041>
- Wilson, H. V. (1907). On Some Phenomena of Coalescence and Regeneration in Sponges. *The Journal of Experimental Zoology*, V(2), 245–258.
- Wojtowicz, W. M., Wu, W., Andre, I., Qian, B., Baker, D., & Zipursky, S. L. (2007). A Vast Repertoire of Dscam Binding Specificities Arises from Modular Interactions of Variable Ig Domains. *Cell*, 130(6), 1134–1145. <https://doi.org/10.1016/j.cell.2007.08.026>
- Wright, S. (1939). The Distribution of Self-Sterility Alleles in Populations. *Genetics*, 24(July), 538–552.
- Wu, Jennifer, Saupe, S. J., & Glass, N. L. (1998). Evidence for balancing selection operating at the het-c heterokaryon incompatibility locus in a group of filamentous fungi. *Proceedings of the National Academy of Science*, 95(October), 12398–12403.
- Wu, Juyou, Wang, S., Gu, Y., Zhang, S., Publicover, S. J., & Franklin-Tong, V. E. (2011). Self-Incompatibility in *Papaver rhoeas* Activates Nonspecific Cation Conductance Permeable to. *Plant Physiology*, 155(February), 963–973. <https://doi.org/10.1104/pp.110.161927>
- Wu, L., Williams, J. S., Wang, N., Khatri, W. A., Roma, D. S., & Kao, T. (2018). Use of Domain-Swapping to Identify Candidate Amino Acids Involved in Differential Interactions between Two Allelic Variants of Type-1 S-Locus F-Box Protein and S3-RNase in *Petunia inflata*. *Plant & Cell Physiology*, 59(November 2017), 234–247. <https://doi.org/10.1093/pcp/pcx176>
- Xing, S., Li, M., & Liu, P. (2013). Evolution of S-domain receptor-like kinases in land plants and origination of S-locus receptor kinases in Brassicaceae. *BMC Evolutionary Biology*, 13(69).
- Yang, A.-S., & Honig, B. (2000). An Integrated Approach to the Analysis and Modeling of Protein Sequences and Structures. III. A Comparative Study of Sequence Conservation in Protein Structural Families using Multiple Structural Alignments. *Journal of Molecular Biology*, 301, 691–711. <https://doi.org/10.1006/jmbi.2000.3975>
- Yang, H., Hu, E., Buchanan, J. T., & Tiersch, T. R. (2018). A Strategy for Sperm Cryopreservation of Atlantic Salmon, *Salmo salar*, for Remote Commercial-scale High-throughput Processing. *Journal of the World Aquaculture Society*, 49(1), 96–112. <https://doi.org/10.1111/jwas.12431>
- Yund, P., & Feldgarden, M. (1992). Rapid Proliferation of Historecognition Alleles in Populations of a Colonial Ascidian. *The Journal of Experimental Zoology*, 263, 442–452.
- Zepeda-Rivera, M. A., Saak, C. C., & Gibbs, K. A. (2018). A Proposed Chaperone of the Bacterial

- Type VI Secretion System Functions To Constrain a Self-Identity Protein. *Journal of Bacteriology*, 200(14), 1–16.
- Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7), 2302–2309. <https://doi.org/10.1093/nar/gki524>
- Zhao, J., Gladieux, P., Hutchison, E., Bueche, J., Hall, C., Perraudau, F., & Glass, N. L. (2015). Identification of Allorecognition Loci in *Neurospora crassa* by Genomics and Evolutionary Approaches. *Molecular Biology and Evolution*, 32(9), 2417–2432. <https://doi.org/10.1093/molbev/msv125>
- Ziegenfuss, J. S., Biswas, R., Avery, M. A., Hong, K., Sheehan, A. E., Yeung, Y.-G., Stanley, E. R., & Freeman, M. R. (2008). Draper-dependent glial phagocytic activity is mediated by Src and Syk family kinase signalling. *Nature*, 453, 935–940. <https://doi.org/10.1038/nature06901>
- Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology*, 430(15), 2237–2243. <https://doi.org/10.1016/j.jmb.2017.12.007>