Genetic Diversity of the Allodeterminant alr2 in Hydractinia symbiolongicarpus

Rafael D. Rosengarten,†1 Maria A. Moreno,1 Fadi G. Lakkis,2 Leo W. Buss,3,4 and Stephen L. Dellaporta*,1

1Department of Molecular, Cellular and Developmental Biology, Yale University
2Thomas E. Starzl Transplantation Institute, Departments of Surgery and Immunology, University of Pittsburgh School of Medicine
3Department of Ecology and Evolutionary Biology, Yale University
4Department of Geology and Geophysics, Yale University
†Present Address: Joint BioEnergy Institute, Emeryville, California.
*Corresponding author: E-mail: stephen.dellaporta@yale.edu.

Associate editor: Yoko Satta

Abstract

Hydractinia symbiolongicarpus, a colonial cnidarian (class Hydrozoa) epibiont on hermit crab shells, is well established as a model for genetic studies of allore cognition. Recently, two linked loci, allore cognition (alr) 1 and alr2, were identified by positional cloning and shown to be major determinants of histocompatibility. Both genes encode putative transmembrane proteins with hypervariable extracellular domains similar to immunoglobulin (Ig)-like domains. We sought to characterize the naturally occurring variation at the alr2 locus and to understand the origins of this molecular diversity. We examined full-length cDNA coding sequences derived from a sample of 21 field-collected colonies, including 18 chosen haphazardly and two laboratory reference strains. Of the 35 alleles recovered from the 18 unbiased samples, 34 encoded unique gene products. We identified two distinct structural classes of alleles that varied over a large central region of the gene but both possessed highly polymorphic extracellular domains I, similar to an Ig-like V-set domain. The discovery of structurally chimeric alleles provided evidence that interallelic recombination may contribute to alr2 variation. Comparisons of the genomic region encompassing alr2 from two field-derived haplotypes and one laboratory reference sequence revealed a history of structural variation at the haplotype level as well. Maintenance of large numbers of equally rare alleles in a natural population is a hallmark of negative frequency-dependent selection and is expected to produce high levels of heterozygosity. The observed alr2 allelic diversity is comparable with that found in immune recognition molecules such as human leukocyte antigens, B cell Igs, or natural killer cell Ig-like receptors.

Key words: histocompatibility, allore cognition, frequency-dependent selection, hypervariability.

Introduction

Allore cognition, the ability to distinguish oneself from genetically distinct conspecifics, is a phenomenon widely observed among benthic colonial invertebrates (Bancroft 1903; Hauenschild 1954, 1956; Ivker 1972; Van de Vyver 1975; Hildemann et al. 1977; Chaney 1983; Grosberg 1988). These sessile individuals grow by clonal propagation and frequently come into contact with adjacent organisms, imposing competition for limited spatial resources (Jackson 1977, 1985). The decision to fuse into a chimera, or reject and fight for substrate space, is critical for mediating competition at the level of the organism. When contacting colonies fuse, individuals then become susceptible to competition at the level of cell lineage. Colonial marine invertebrates do not sequester their germ lines as embryos, rather they maintain a population of multipotent stem cells that give rise to gametes later in ontogeny (Buss 1983, 1987). In a chimeric colony, the stem cells of one individual could outcompete those of other, effectively parasitizing the latter’s germ line. Thus, allore cognition determines whether colonies compete for space or whether cell lineages compete for gametic output.

Hydractinia symbiolongicarpus is a colonial cnidarian (class Hydrozoa) that for decades has been a model for studies on allore cognition (Hauenschild 1954, 1956; Ivker 1967, 1972; Grosberg and Quinn 1986; Lange et al. 1989; Buss and Grosberg 1990; Buss and Shenk 1990; Shenk and Buss 1991; Grosberg et al. 1996; Mokady and Buss 1996; Cadavid et al. 2004). Hydractinia is one of two invertebrate systems (the other is the chordate Botryllus schlosseri) with well-characterized allore cognition genetics. Given the ancient divergence of the cnidarian lineage, studies in this organism allow consideration of metazoan immunity over vast evolutionary time.

Hydractinia colonies typically grow on hermit crab shells (Buss and Yund 1989), and when two (or more) individuals come into contact, they undergo a fusion/rejection allore cognition response (Buss 1990). Fusion occurs infrequently, between 1.8% and 4.4% of field-collected colony pairs (Grosberg et al. 1996; Mokady and Buss 1996; Nicotra and Buss 2005). The fusion process is controlled by a cell–cell recognition event manifested as ectodermal cell adhesion and subsequent establishment of a contiguous endodermal gastrovascular system (Buss et al. 1984; Lange et al.
In contrast, the more frequent outcome is rejection, which occurs between wild-type colony pairs at an average rate of 97% (Nicotra and Buss 2005). In rejection responses, ectodermal cells fail to adhere, nematocytes are extensively recruited along the contact area, and nematocytes, harpoon-like organelles, are discharged to destroy the adjacent tissue (Buss et al. 1984; Lange et al. 1989). Colonies may also display numerous versions of transitory fusion, in which they initially fuse and then later separate (Hauenschild 1954, 1956; Müller 1964; Ivker 1972; Grosberg et al. 1996; Cadavid et al. 2004; Powell et al. 2007). The rarity of fusion implies a high degree of diversity in the genetic control of allorecognition.

Classical breeding experiments have shown fusibility in inbred laboratory lines is determined by two linked codominant loci, termed allorecognition (alr) 1 and 2, that reside in a chromosomal region known as the allorecognition complex (ARC) (Mokady and Buss 1996; Cadavid et al. 2004; Powell et al. 2007). Both alr1 and alr2 have been isolated by positional cloning and are predicted to encode proteins with similar domain architectures (Nicotra et al. 2009; Rosa et al. 2010). ALR1 consists of two hypervariable extracellular domains similar to Ig-like domains, a transmembrane domain, and a cytoplasmic tail with a putative immunoreceptor tyrosine–based activation motif (ITAM) (Rosa et al. 2010). The alr2 gene, meanwhile, consists of nine exons and produces a ~2.1 kb transcript, predicted to encode a transmembrane protein with three extracellular domains similar to Ig-like domains and an intracellular domain containing an immunoreceptor tyrosine–based inhibition motif (ITIM) (Nicotra et al. 2009). The display of multiple noncanonical Ig-like domains suggests that both ALR1 and ALR2 may be involved in protein–protein interactions at the cell surface, while the presence of putative cytoplasmic motifs indicates that these molecules may be capable of downstream signaling.

Both defined alr genes are also similar in that the genomic regions encompassing these loci are structurally complex. The 700-kb interval flanking alr1 contains numerous additional open reading frames (ORFs) with strong sequence similarity to the tandem Ig-like domains found in the functional locus, raising the possibility that recombination occurs between members of this alr1-related subcomplex. At alr2, two pseudogenes (herein alr2P1 and alr2P2) were identified adjacent to the functional locus, the products of partial duplications that spanned the 5′ half of the gene, including all three extracellular domains (Nicotra et al. 2009). Together, alr2 and its adjacent pseudogenes comprise an alr2-related region within the ARC.

Initial comparisons of the alr2 gene, including 14 alleles from this study, demonstrated the hypervariability of domain I (Nicotra et al. 2009). Sixty-eight of 111 amino acid positions in domain I varied between at least two alleles. Field-collected colony LH06:082 (formerly LH82) displayed transitory fusion with a laboratory tester colony homozygous for the alr2-F allele and expressed an alr2 allele with a domain I identical to that of alr2-F. Fusibility between these colonies implied that domain I confers recognition specificity or alr2 allototype (Nicotra et al. 2009). Two additional colony pairs (LH06:058 and LH07:014; LH06:003 and LH06:049) have been reported to harbor matching alr2 domains I, as well as matching extracellular domains at alr1 (Rosa et al. 2010). These colony pairs display fusion or transitory fusion, consistent with the hypothesis that domain I of alr2 contributes to allotype.

Alr2 is a prototypical allorecognition determinant. Its structural characteristics and hypervariable extracellular domain similar to an Ig-like V-set are in line with expectations based on numerous examples of other recognition system molecules. Notably, remnants models include genes involved in both adaptive and innate immunity in vertebrates, as well as participants in invertebrate histocompatibility (Litman et al. 2003; Litman et al. 2007). Despite lacking some features of canonical Ig superfamilies, members, such as highly conserved cysteine residues, the observed sequence and predicted structural similarity of the alr2 extracellular domains to Ig-like V-set and I-set domains (Nicotra et al. 2009) invites comparisons with vertebrate B cell receptors, and classical Ig antibodies. These molecules display a pattern of sequence polymorphism especially localized to the extracellular variable (V) region, introduced by a process called somatic hypermutation (SHM) (Steele et al. 1992; Kalberg et al. 1996; Denepoux et al. 1997; Odegaard and Schatz 2006; Di Noia and Neuberger 2007). Similar to alr2, variation between alleles of the most diverse major histocompatibility complex (MHC) genes tends to be confined to the extracellular domain-encoding exons 2 and 3. Human leukocyte antigen (HLA) diversity results from balancing selection acting on variation that arises by point mutations and/or intergenic gene conversion (Takahata and Satta 1998). HLA class II molecules in particular display a signature of frequent recombination—a large suite of novel alleles with distinct combinations of shared polymorphisms (Gorski and Mach 1986; Mach et al. 1986; Huang et al. 1995; von Salome and Kukkonen 2008). Yet another comparison may be made between alr2 and killer-cell immunoglobulin–like receptors (KIRs). These are polymorphic molecules that recognize MHC class I cell surface moieties and often contain inhibitory motifs (i.e., ITIMs) in their cytoplasmic domains (Parham 2005).

In this paper, an extraordinary degree of alr2 allele diversity was assessed from a field survey of full-length cDNA coding sequences (CDs). Nearly, all sampled alleles encoded unique gene products, with sequence variation largely concentrated over the outermost extracellular domain I. Variation over the 5′ end of alr2 alleles sequenced from genomic DNA (gDNA) was found to begin ~150 bp from the start of transcription and increased sharply over exon 2, which encodes domain I. In addition to sequence variation, alleles were found to belong to one of two classes based on substantial structural polymorphism. We further compared the genomic interval spanning the alr2-related genes from two field-derived haplotypes and the reference sequence alr2-F, data that indicated a rich history of duplications, insertions and deletions, and rearrangements that may have contributed to alr2 diversity at the level of haplotype. The characterization of sequence...
and structure-based polymorphisms that contribute to allelic novelty strongly implicated negative frequency-dependent selection as an evolutionary force maintaining allele variation.

Materials and Methods
All polymerase chain reaction (PCR) primer sequences and specifications are provided in supplementary table 3 (Supplementary Material online). Custom Perl scripts used in sequence assembly, annotation, and analysis are available upon request.

Animal Collection and Maintenance
Hermit crab shells bearing large Hydractinia colonies were collected along the Connecticut coast in 2006 and 2007. All but three colonies in this study derived from the intertidal at Lighthouse Point (New Haven). Two came from Meig’s Point, Hammonasset State Park, (Madison), and one from a 2002 collection in the shallow subtidal at Old Quarry Harbor (Guilford). Given the close proximity of these locations, and the preponderance of colonies from Lighthouse Point, we herein treat this as a single population. Colonies were explanted from host shells to glass microscope slides and maintained in laboratory seawater aquaria using standard methods (Shenk and Buss 1991).

Nucleic Acid Extraction and cDNA Synthesis
Total nucleic acids were extracted from selected colonies using a modified urea extraction protocol (Chen and Dellaporta 1994). RNA was isolated from total nucleic acid preparations by treating with recombinant RNase-free DNase I (Roche), phenolchloroform extraction, and precipitation in 100% ethanol. Approximately 2 μg of total RNA was used as a template in first-strand cDNA synthesis reaction using the Generacer Superscript III RT module (Invitrogen), following manufacturer’s instructions, with one modification—in addition to the oligo-dT primer, 3μg random hexamer oligos were added to facilitate synthesis of long transcripts.

Amplification, Cloning, and Sequencing of \textit{alr2} Alleles
Full-length CDSs were amplified from first-strand cDNA by reverse transcriptase (RT-) PCR as overlapping 5’ and 3’ segments using primer combinations P2859/P2199 and P2051/P2049. For difficult templates, amplicons were diluted and used to template a nested PCR reaction using the same exon-specific primer and a nested 5’ trans-spliced leader primer (P2882) or nested 3’ untranslated region (UTR) primer (P2663). Amplicons were purified using PEG-NaCl (Lis and Schleif 1975) and cloned with the Stratagene Blunt PCR Cloning Kit (vector pSC-B), following manufacturer’s instructions. Plasmid DNA was prepared from four clones per reaction using an alkaline lysis mini-prep (Voskuil and Chambless 1993) and sequenced in both directions by Sanger sequencing. Partial 5’ gDNA alleles were amplified in two steps: long-range PCR using primers specific to the \textit{alr2} promoter region (P3426) and exon 9 (P2049) and a nested reaction using P3548 and intron 2 primer P3549. Amplicons were purified as above and cloned by ligation-independent cloning (LIC) (Aslanidis and de Jong 1990) into custom LIC-adapted plasmid (pYU2418). Transformants were screened by PCR, and these amplicons sequenced in both directions with vector primers by Sanger sequencing.

Assembly and Annotation
Sequencing reads were assembled using PhredPhrap-Consed v.17 (www.phrap.org). cDNA contigs were annotated with SeqBuilder module in the Lasergene7 software package (DNASTAR). To define exon borders, cDNA sequences for which corresponding gDNA contigs were available were mapped to the gDNA using the NCBI program Spidey (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey). Conserved exon/intron borders were assumed for all other cDNAs.

Analysis of Variation and Selection
Amino acid translations of complete \textit{alr2} CDSs were aligned using Prank \_F (Loytynoja and Goldman 2008) with default settings, and the alignment edited manually. The Shannon diversity index, $H’ = -\Sigma p \times \ln(p)$ (Shannon 1949; Cowell et al. 1998), where $p$ is the proportion of a given residue among all residues at that site, was calculated to represent sequence variability. For analysis of site-wise selection, the amino acid alignment positions were back-translated to the corresponding CDSs. Codons under selection were determined by the Single Likelihood Ancestor Counting, Fixed Effects Likelihood, and Random Effects Likelihood algorithms implemented in the HyPhy software package (Pond and Frost 2005). Sites predicted to be under diversifying selection by at least two of the three models at significance values of $P < 0.05$ and Bayes factor $>50$ were considered to be conserved sites. Calculations of per-domain and per-gene synonymous and nonsynonymous polymorphism were performed using DNAsp v.5 (Librado and Rozas 2009).

Haplotype Cloning and Sequencing
A fosmid library representing 3-5X genomic coverage of colony LH06:003 was constructed from high molecular weight gDNA, using the Epicentre Copy Control Fosmid kit as instructed. Fosmid clones positive for markers in the \textit{alr2} region were sheared, shotgun cloned, and sequenced by Sanger sequencing. Chromatograms were assembled using PhredPhrapConsed, and remaining gaps in the assembly closed by sequencing gap-spanning PCR amplicons. The final raw assembled LH06:003A contig was 105,415 bp long and comprised 610 reads. Assuming ~500 high-quality (Phrap scores >30) base pairs per read, this equates to an average read depth of ~3×, with 75% of the positions ≥2× coverage. The raw LH06:003B contig was 96,031 bp long and comprised 656 reads, equating to 3.4× average read depth. Over 85% of the positions were covered ≥2×.

Haplotype Annotation and Comparisons
Complete haplotype contigs were annotated in the Lasergene7 SeqBuilder module. The exon/intron structure of...
alr2 and related pseudogenes was mapped using Spidey. The presence or absence of additional genes was assessed by comparisons to the alr2- and -f sequences (Nicotra et al. 2009), and by Blastx (Altschul et al. 1997) searches against the NCBI nonredundant protein database, queried with overlapping 1,000 bp segments of each haplotype. Furthermore, ORFs > 300 bp were predicted by SeqBuilder and identities assigned by Blastx analysis. Given the high degree of overall similarity to the alr2 and -f reference sequences, which had been thoroughly analyzed for gene coding potential (Nicotra 2007), ab initio gene prediction was not performed on the LH06:003 haplotypes. Contigs were trimmed to begin at the start codon of the atp synthase gene and to end at the 5’ UTR of the fucosyl transferase (FUT) gene on the opposite strand. Comparisons between haplotypes were made using Pipmaker (Schwartz et al. 2000) and MultiPipMaker (Schwartz et al. 2003). Repetitive DNA tracts were identified with RepeatModeler (Smit A, Hubley R, Green P, unpublished data) and confirmed by sequence similarity searches.

Results

Diversity of alr2 Alleles

Eighteen field-collected colonies were chosen haphazardly for analysis of full-length CDSs. Thirty-five of 36 possible alleles were recovered from these colonies, with only one individual (LH07:043) failing to yield two distinct alr2 alleles—thus, 17 of 18 colonies were heterozygous at alr2 (table 1). Of these 35 alleles, 34 encoded unique gene products and 29 encoded unique extracellular domains, thought to be involved in recognition specificity. No colony was found to express more than two alleles, thus we did not find evidence that variants arise somatically. Three additional field-collected colonies were sampled based on pre-screening for matching domain I sequences (Rosengarten 2010) or known fusibility phenotypes (Nicotra et al. 2009). The alr2- and alr2- laboratory reference alleles were also sequenced, cloned from colony 14-7 (Mokady and Buss 1996). In total, 43 alr2 alleles were sequenced. Fourteen of the wild-type alleles characterized in this study were previously reported upon the identification of alr2 (Nicotra et al. 2009). The animal source, GenBank Accessions, and nomenclature conventions for all alleles are provided in supplementary table 1 (Supplementary Material online).

Each CDS was compared by Blast2 (Altschul et al. 1997) to the alr2- and -f reference sequences. Twenty-four of the 41 wild-type alleles readily aligned to the alr2- and -f references and were designated “type I.” The remaining 17 alleles showed significant dissimilarity to type I alleles in a central section comprised exons 5–7 and were designated “type II.” Further details of this structural polymorphism are presented below and in the supplementary Results and Discussion (Supplementary Material online). Of the 20 field-derived colonies for which both alr2 alleles were cloned, eight colonies were found to express only type I alleles, four colonies were found to express only type II alleles, and eight colonies were found to express an allele of both type (supplementary table 1, Supplementary Material online). These results suggested that type I and II genes are allelic, an interpretation that was later confirmed by haplotype sequencing (discussed below), and occur at similar frequencies in the sampled population. For subsequent sequence comparisons, type I and type II alleles were partitioned into their respective groups.

For the unbiased subset of alleles (n = 35), all pairwise comparisons were performed by the FastA program ssearch (Pearson 1996), and percent amino acid identities calculated within each group of alleles (fig. 1A). Of the 301 possible allele pairs (210 type I, 91 type II), only six pairs of predicted proteins (2%) had identical extracellular domains, a value consistent with the 1.8–4.4% of wild-type pairs reported to fuse (Nicotra and Buss 2005). Type I and II alleles showed greater amino acid identity and similarity within a class than between classes. The distribution of pairwise amino acid identity of the type II group was skewed toward lower percent identity relative to the type I distribution (fig. 1B). Mean pairwise amino acid identity was 93% among type I and 91% among type II alleles (student’s t-test with unequal variance, t = 2.7; P < 0.05; degrees of freedom [df] = 300).

Structural Similarity and Variability

A multiple sequence alignment (MSA) of the predicted protein sequences from the complete set of 41 field-derived alleles and the reference alleles alr2-f and -r was performed using Prank+ (Loytynoja and Goldman 2008) and manually edited for consistency in gap placement (supplementary Alignment File, Supplementary Material online). Four of the alleles contained an alternative start codon that extended the ORF by ten or 12 codons. In the case of the 12 codon extensions, the upstream ATG was the only possible start codon, whereas in the ten codon extensions, the canonical ATG found in most alleles was conserved as well. Alleles ranged in length from 1,962 to 2,049 bases, encoding proteins from 654 to 683 amino acids in size. All 43 alleles shared an N-terminus signal peptide motif predicted by SignalP 3.0 (Emanuelsson et al. 2007) with the conserved cleavage motif VS[C/S] × LS (position 30–31 in the MSA, supplementary Alignment File, Supplementary Material online). Other common features included a highly variable domain with predicted structural similarities to an Ig-like V-set, followed by two less variable domains predicted to be similar to Ig-like I-sets, an extracellular “spacer” region of no homology to known sequences, a single transmembrane-spanning domain, and related pseudogenes was mapped using Spidey. The presence or absence of additional genes was assessed by comparisons to the alr2- and -f sequences (Nicotra et al. 2009), and by Blastx (Altschul et al. 1997) searches against the NCBI nonredundant protein database, queried with overlapping 1,000 bp segments of each haplotype. Furthermore, ORFs > 300 bp were predicted by SeqBuilder and identities assigned by Blastx analysis. Given the high degree of overall similarity to the alr2 and -f reference sequences, which had been thoroughly analyzed for gene coding potential (Nicotra 2007), ab initio gene prediction was not performed on the LH06:003 haplotypes. Contigs were trimmed to begin at the start codon of the atp synthase gene and to end at the 5’ UTR of the fucosyl transferase (FUT) gene on the opposite strand. Comparisons between haplotypes were made using Pipmaker (Schwartz et al. 2000) and MultiPipMaker (Schwartz et al. 2003). Repetitive DNA tracts were identified with RepeatModeler (Smit A, Hubley R, Green P, unpublished data) and confirmed by sequence similarity searches.

Results

Diversity of alr2 Alleles

Eighteen field-collected colonies were chosen haphazardly for analysis of full-length CDSs. Thirty-five of 36 possible alleles were recovered from these colonies, with only one individual (LH07:043) failing to yield two distinct alr2 alleles—thus, 17 of 18 colonies were heterozygous at alr2 (table 1). Of these 35 alleles, 34 encoded unique gene products and 29 encoded unique extracellular domains, thought to be involved in recognition specificity. No colony was found to express more than two alleles, thus we did not find evidence that variants arise somatically. Three additional field-collected colonies were sampled based on pre-screening for matching domain I sequences (Rosengarten 2010) or known fusibility phenotypes (Nicotra et al. 2009). The alr2- and alr2- laboratory reference alleles were also sequenced, cloned from colony 14-7 (Mokady and Buss 1996). In total, 43 alr2 alleles were sequenced. Fourteen of the wild-type alleles characterized in this study were previously reported upon the identification of alr2 (Nicotra et al. 2009). The animal source, GenBank Accessions, and nomenclature conventions for all alleles are provided in supplementary table 1 (Supplementary Material online).

Each CDS was compared by Blast2 (Altschul et al. 1997) to the alr2- and -f reference sequences. Twenty-four of the 41 wild-type alleles readily aligned to the alr2- and -f references and were designated “type I.” The remaining 17 alleles showed significant dissimilarity to type I alleles in a central section comprised exons 5–7 and were designated “type II.” Further details of this structural polymorphism are presented below and in the supplementary Results and Discussion (Supplementary Material online). Of the 20 field-derived colonies for which both alr2 alleles were cloned, eight colonies were found to express only type I alleles, four colonies were found to express only type II alleles, and eight colonies were found to express an allele of both type (supplementary table 1, Supplementary Material online). These results suggested that type I and II genes are allelic, an interpretation that was later confirmed by haplotype sequencing (discussed below), and occur at similar frequencies in the sampled population. For subsequent sequence comparisons, type I and type II alleles were partitioned into their respective groups.

For the unbiased subset of alleles (n = 35), all pairwise comparisons were performed by the FastA program ssearch (Pearson 1996), and percent amino acid identities calculated within each group of alleles (fig. 1A). Of the 301 possible allele pairs (210 type I, 91 type II), only six pairs of predicted proteins (2%) had identical extracellular domains, a value consistent with the 1.8–4.4% of wild-type pairs reported to fuse (Nicotra and Buss 2005). Type I and II alleles showed greater amino acid identity and similarity within a class than between classes. The distribution of pairwise amino acid identity of the type II group was skewed toward lower percent identity relative to the type I distribution (fig. 1B). Mean pairwise amino acid identity was 93% among type I and 91% among type II alleles (student’s t-test with unequal variance, t = 2.7; P < 0.05; degrees of freedom [df] = 300).

Structural Similarity and Variability

A multiple sequence alignment (MSA) of the predicted protein sequences from the complete set of 41 field-derived alleles and the reference alleles alr2-f and -r was performed using Prank+ (Loytynoja and Goldman 2008) and manually edited for consistency in gap placement (supplementary Alignment File, Supplementary Material online). Four of the alleles contained an alternative start codon that extended the ORF by ten or 12 codons. In the case of the 12 codon extensions, the upstream ATG was the only possible start codon, whereas in the ten codon extensions, the canonical ATG found in most alleles was conserved as well. Alleles ranged in length from 1,962 to 2,049 bases, encoding proteins from 654 to 683 amino acids in size. All 43 alleles shared an N-terminus signal peptide motif predicted by SignalP 3.0 (Emanuelsson et al. 2007) with the conserved cleavage motif VS[C/S] × LS (position 30–31 in the MSA, supplementary Alignment File, Supplementary Material online). Other common features included a highly variable domain with predicted structural similarities to an Ig-like V-set, followed by two less variable domains predicted to be similar to Ig-like I-sets, an extracellular “spacer” region of no homology to known sequences, a single transmembrane-spanning domain.
and a conserved cytoplasmic domain, all consistent with the structure first identified by Nicotra et al. (2009).

Per-site variability ($H'$) was calculated from the amino acid MSA and mapped to the alr2 coding sequence (fig. 2A). As previously described, alleles were categorized based on structural similarity (type I) or dissimilarity (type II) to the alr2-f and -r reference alleles. The greatest dissimilarity between types I and II was found in a central section encoded by exons 5–7. This region included the extracellular “spacer,” transmembrane domain, and beginning of the cytoplasmic domain. Inspection of the ALR2 alignment identified type I- and II-specific polymorphisms—indels and substitutions—that flanked the central region (fig. 2A, supplementary fig. 1, Supplementary Material online). The diagnostic features began within the exon 3-encoded domain II and were dispersed throughout the exon 4-encoded domain III. On the 3’ side of the central section, three type-specific substitutions were observed near the junctions of exons 8 and 9, despite an otherwise high degree of conservation in this region. Although the flanking variation was not as extensive as the dimorphic central region, the type-specific polymorphisms in domains II and III and in the cytodomain provided further support for the classification of alr2 alleles into two distinct groups.

Chimeric alleles possessing features of both types I and II were also found (fig. 2B, supplementary fig. 1, Supplementary Material online). Four alleles (LH06:008a, LH06:053b, LH07:014b, and LH07:041a) contained the central and C-terminal region (exons 5–9) that defined these alleles as type I (red box in supplementary fig. 1, Supplementary Material online). Nevertheless, type II-specific features were identified in domains II and III of these sequences. Three other alleles (LH06:003b, LH07:026b, and LH07:060b) shared all indels and most residues of type II sequences but contained a four amino acid tract in domain III that was specific to type I (fig. 2B, yellow box in supplementary fig. 1, Supplementary Material online). Two other possible cases of chimerism were seen between alleles of different structure types but with closely matching domains I (fig. 2C). Allele LH06:028a, which clearly belonged to the type II structure group as defined by exons 3–9, possessed a domain I nearly identical (differing by only 1 amino acid) to that of three type I alleles—alr2-r, LH06:003a, and LH06:049a. Likewise, domain I of LH06:053b and LH06:058b differed by only one residue, but the alleles varied in structure type. These alleles indicated that recombination may contribute to alr2 diversity. The ancestries of the different structure types are further considered in the supplementary Results and Discussion (see also supplementary fig. 3, Supplementary Material online).

**Sequence Variation of alr2 cDNAs**

Per-site variability ($H'$) from type-specific amino acid MSAs was plotted over the length of the predicted proteins (fig. 3). The preponderance of $H'$ peaks were found in domain I (V-set-like) of both type I ($n = 26$) and type II ($n = 17$) alleles (see also supplementary fig. 4, Supplementary Material online). Variable sites were also seen in domains II and III (I-set-like) and scattered elsewhere in the alignment. Domain III and the extracellular spacer were more variable among type II alleles than their type I counterparts (student’s t-test with unequal variance, $t = 0.68, 0.34; P < 0.05; df = 190, 179$). Nevertheless, the exon 2–encoded domain I was at least twice as variable, on average, as any other domain. Seventy-eight of 117 amino acid alignment positions encoded by exon 2 were variable in at least one allele.

The observed pattern of variation could arise from either neutral mutation with strong selection, elevated mutation under neutral selection, or elevated mutation coupled with strong selection. We tested for a signature of selection based on the per-site ratio of nonsynonymous ($dN$) to
synonymous (\(dS\)) substitutions. Nine codons were identified with significantly elevated \(dN/dS\) and were inferred to be under diversifying selection (red bars in fig. 3, supplementary table 2, Supplementary Material online). The majority of these codons were found in exon 2, consistent with the analysis of selection reported by Nicotra et al. (2009). We then asked if this selection could alone account for variation in the extracellular domains or if mutations occurred in these regions at an elevated rate. The synonymous (\(\pi(s)\)) and nonsynonymous (\(\pi(a)\)) substitution rates were determined for each domain-coding region and for the entire \(alr2\) coding sequence, from type I and II alignments (fig. 4). Silent polymorphisms in most extracellular domains were observed at rates elevated over that of the entire gene, whereas \(\pi(s)\) in the central and cytoplasmic regions were below the entire gene \(\pi(s)\). The mean \(\pi(s)\) of the hypervariable domain-encoding exon 2, in particular, was 2.6 (type I) and 1.8 (type II) fold higher than the all-gene level (student’s \(t\)-test with unequal variance, \(t = 384, 199; P << 0.05; df = 324, 135\)). Consistent with the amino acid

![Fig. 2. Structural polymorphisms defined two classes of \(alr2\) alleles, type I and type II, and provided evidence for interallelic recombination. (A) Per-site sequence variation (\(H'\)) is displayed as a heat map over the \(alr2\) exon structure. \(H' = -\sum p \ln(p)\), where \(p\) is the proportion of a given residue at each site, was calculated from an amino acid MSA of all alleles (merged, \(n = 43\)) and from MSAs partitioned by allele type (type I, \(n = 26\); type II, \(n = 17\)). High levels of variation were observed over exon 2, which encodes the hypervariable domain I, as expected. Extensive structural variation in a central region spanning exons 5–7 (solid red line) defined two structural classes, types I and II. Within each type, this region is largely conserved but is divergent between types. Type-specific polymorphisms, that is, substitutions and codon indels, were observed flanking the central region from exon 3 to exon 9 (dotted red line). SP = signal peptide, ECS = extracellular spacer region, TM = transmembrane spanning domain. (B) Two sets of alleles were identified possessing features of both type I (violet shading) and type II (green shading). In the first instance, alleles displayed type I structural characteristics but type II flanking polymorphisms. A second group was observed with a patch of type I sequence within otherwise type II features. This second group contained an additional allele-specific sequence tract (orange bar). (C) Two sets of alleles were identified possessing nearly identical domain I sequences (cyan and yellow boxes) but different structure types.

under diversifying selection. SP alleles (extracellular spacer region, TM) main I exhibited a highest in the hypervariable region. Only the V-set-like domain I suggested that many but not all variable sites and those under diversifying selection may reside in putative loops (data not shown). Without a solved structure, however, the exact location of variable sites remains inconclusive.

We sought to further characterize the alr2 sequence variability to determine where in the gene variability begins and whether it persists over exon–intron borders or is restricted to coding sequence. This analysis required gDNA contigs that spanned the start of transcription and intronic sequence in order to assess the distance of variable sites from the start of transcription. Twenty-four contigs, which included the first ~1.4–1.8 kb of alr2 gDNA, were amplified and sequenced from 19 of the H. symbiolongicarpus field isolates (supplementary table 1, Supplementary Material online). The sequences were aligned with the alr2-f and -r reference alleles, beginning 29 bp before the start of the 5’ UTR and continuing into intron 2. Per-site variation (H’) was calculated and plotted over the length of the nucleotide alignment (supplementary fig. 2, Supplementary Material online). H’ was low at the beginning of the alignment and began to increase between position 100 and 150, roughly corresponding to the start of intron 1. The variability peaked in exon 2 around position 375, then declined until the beginning of intron 2 (position 635). H’ then increased dramatically over intron 2. This increase was due to both substantial structural variation (indels) in the intron and point mutations between alleles. The overall H’ distribution was consistent with π(s) calculated for exons 1 and 2 and the noncoding regions (supplementary fig. 2, Supplementary Material online).

Comparative Analysis of alr2 Haplotypes
In order to examine the contribution of haplotype variation to alr2 diversity, we sequenced the genomic region spanning alr2 and the linked pseudogenes from both haplotypes of field-isolate LH06:003. These genomic intervals, or alr2 haplotypes, were designated as alr2-LH06:003A and B (GenBank accession HM017511 and HM017512, respectively).

The size, gene content, and gene organization of four alr2 haplotypes (previously sequenced laboratory reference haplotypes alr2-F and -R and newly sequenced wild-type LH06:003A and B) were compared with one another. Sizes of these syntenic regions ranged from 93,737 bp (LH06:003B) to 118,348 bp (alr2-F), a difference of 24,611 bp. LH06:003A and alr2-R were nearly identical in size and sequence (see supplementary Results and Discussion, Supplementary Material online), which was expected, as LH06:003A was previously found to be highly similar to ARC-R at numerous genotyping markers and to express a nearly identical alr2 allele (Nicotra 2007). Because further analyses focus on alr2 and the alr2-related duplications, which are nearly invariant between A and R, results are reported for the LH06:003A sequence only.

The alr2 haplotypes were defined by the synteny of flanking genes atp synthase and FUT (fig. 5, supplementary fig. 5, Supplementary Material online). Internally, the A and
B haplotypes each contained the genes previously identified for \textit{alr2-F} and \textit{-R} (Nicota et al. 2009). This included a full-length \textit{alr2} locus as well as two partial duplications (formerly CDS6P and CDS5P) herein referred to as \textit{alr2P1} and \textit{alr2P2}. A conserved ORF of unknown function (d on fig. 5A, supplementary fig. 5, Supplementary Material online) resided between \textit{alr2} and \textit{FUT}. In our analysis, the \textit{alr2-F} haplotype was found to contain an additional partial duplication of \textit{alr2}-related sequence (\textit{alr2P3}) immediately upstream of \textit{alr2P2}. This duplication was most similar to a 647-bp interval of \textit{alr2P2} (76\% identical, \textit{e} value = 4 × 10^{-133}) beginning 166 bp prior to the 5' UTR and spanning 214 bp into exon 2. Together \textit{alr2} and its related duplications constitute an \textit{alr2} subcomplex within the larger \text{ARC}.

Numerous ORFs were identified with significant sequence similarity to mobile elements such as tyrosine recombinase \textit{(a)}, reverse transcriptase \textit{(b,c)}, copia-type retrotransposon \textit{(e)}, and tigger transposable element–derived protein \textit{(g)} (fig. 5A, supplementary fig. 5, Supplementary Material online). Some of these were haplotype specific, such as elements \textit{(e)} found only on the LH06:003A haplotype or \textit{(g)} on the B haplotype. Others, such as the tandemly arranged tyrosine recombinase \textit{(a)} and reverse transcriptase \textit{(b)}, were shared between two haplotypes. Two sequences were found on LH06:003A with high similarity \textit{(95\% identity)} to each other and to the reverse transcriptase \textit{(c)} on the \textit{F} haplotype. The single \textit{(c)} ORF on \textit{alr2-F} was found in the opposite 5'–3' orientation as the two \textit{(c)} sequences on LH06:003A. In addition to these ORFs, numerous repetitive tracts were identified in intergenic regions, some of which bore significant similarity to segments of introns 2 and 3 (fig. 5B, see also supplementary Results and Discussion, Supplementary Material online). These mobile elements and repetitive tracts could possibly contribute to structural heterogeneity of the \textit{alr2} genomic region.

\textbf{Alr2 and alr2-Related Pseudogenes}  
The structures of \textit{alr2}, \textit{alr2P1}, and \textit{alr2P2} were examined by aligning \textit{alr2} cDNA sequences with corresponding genomic regions. The \textit{alr2} genes on the haplotypes \textbf{A}, \textbf{B}, and \textbf{F} comprised nine exons (supplementary fig. 6, Supplementary Material online). On \textbf{A} and \textbf{F}, these nine exons spanned 16.1 and 15.8 kb, respectively, and encoded type I \textit{alr2} transcripts. LH06:003B contained a type II \textit{alr2} gene that spanned only 8.5 kb. This dramatically smaller gene size was due primarily to shortened introns 2, 3, 4, and 8.

Next, the composition of \textit{alr2P1} and \textit{alr2P2} were analyzed (supplementary fig. 6, Supplementary Material online). The duplicated regions all began upstream of the \textit{alr2} promoter and 5' UTR. On haplotypes \textbf{B} and \textbf{F}, \textit{alr2P1} and \textit{alr2P2} were predicted to extend as far as the first 5 exons of \textit{alr2}. Haplotype \textbf{A} possessed a similar five exon \textit{alr2P2}, but only exons 1–4 were present in \textit{alr2P1}. All \textit{alr2P1} and \textit{alr2P2} genes displayed type II polymorphisms in exons 3 and 4 and, when present, a distinct type II exon 5 (supplementary fig. 6, Supplementary Material online).

\begin{figure}[ht]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Per-domain nucleotide diversity. Rates of synonymous (\(\pi(s)\)) and nonsynonymous (\(\pi(a)\)) polymorphism were determined for each domain-encoding region (data points) for type I (A) and type II (B) alleles. The solid black line indicates the \(\pi(s)\) value calculated for the entire gene. The ratio of \(\pi(a)\) to \(\pi(s)\) is indicated by X's. The dotted line indicates \(\pi(a)/\pi(s) = 1\). SP = signal peptide, ECS = extracellular spacer region, TM = transmembrane domain.}
\end{figure}
The type II nature of these genes was observed regardless of the allele type of the adjacent $alr2$ gene. On the A and F haplotypes, for instance, $alr2P1$ and $alr2P2$ were both type II-like but $alr2$ was, by definition, a type I allele. The introns of the $alr2P$ genes were also similar in size and sequence to that of the type II $alr2$-$LH06:003b$ allele. Together, these

**Fig. 5.** Circos display of the genomic interval spanning $alr2$ from two wild-type haplotypes, $LH06:003A$ and $LH06:003B$, and laboratory reference haplotype $alr2-F$. Ribbons, repeats, and gene features are color coded as indicated on the figure legend. (A) Genomic contigs delineated by non-$alr2$–related syntenic genes $atp$ synthase ($atp$) and $FUT$. $Alr2$ and related pseudogenes ($P#$) are labeled. Additional features include ORFs of unknown function ($d, f$), and putative transposable elements ($a, b, c, e, g$). (B) Repetitive sequence tracts identified by RepeatModeler and BlastN searches. (C) Regions of intrachromosomal similarity, identified by PipMaker. (D) Syntenic gene positions and regions of interchromosomal similarity, identified by MultiPipMaker.
data suggested that the alr2P sequences are more related to type II than to type I alr2 alleles.

Signatures of recombination were analyzed with DNAsp v.5 (Librado and Rozas 2009) by partitioning the alr2 and alr2P sequences from the A, B, and F haplotypes into two groups for comparison. In one instance, subset 1 contained all three alr2 loci and subset 2 the alr2P genes. Alternatively, sequences were grouped such that all three genes from one haplotype were compared with the genes from the other two haplotypes. Two short tracts (nine and four nucleotides, respectively) were identified in either alr2P gene on LH06:003A, depending on how the sequences were grouped (data not shown). These results were inconsistent with the expectation of gene conversion in which pseudogenes act as donors and functional genes as recipients. Given the unknown relationship of these haplotypes, actual donors and recipients cannot be deduced, and these sequence tracts could just as likely have occurred from reciprocal recombination.

Structural Heterogeneity
The extent of the duplicated regions of the alr2 subcomplex was investigated by dot plot analysis and found to be more contiguous on the B haplotype than on A or F (fig. 5C, supplementary fig. 7, Supplementary Material online). The 3′ limit of similarity between the alr2P and alr2-LH06:003b allele was found in intron 5 (discussed above). Nearly, 14 kb of contiguous sequence upstream of alr2 was significantly similar to the corresponding interval upstream of alr2P1. The region of similarity encompassing alr2P2, meanwhile, extended from the 5′ promoter to 3 kb downstream of the gene and was most similar to the corresponding interval around alr2P1. Similar patterns were observed for LH06:003A and alr2-F, although discontinuity precluded precise definition of the duplication boundaries. These results suggest that alr2P1 may have originated as part of a large duplication encompassing intragenic sequence and the 5′ half of alr2, whereas alr2P2 arose as an independent smaller duplication.

Structural variation between haplotypes was also explored by multiPIP analysis, which indicated that haplotype-specific insertion/deletion, duplication, and/or transposition events have taken place during the evolution of these haplotypes. The majority of the size differences, for example, between the largest haplotype, alr2-F, and the others were accounted for by intergenic indels (figs. 5D and 6). Among all haplotypes, intergenic indels ranged in size from a single base pair to a nearly 10 kb sequence present on LH06:003A but absent on LH06:003B. Approximately 5 kb of this large indel was absent on the alr2-F haplotype as well. The majority of intragenic indels, meanwhile, were concentrated in the introns of alr2, alr2P1, and alr2P2 (figs. 5D and 6, supplementary fig. 6, Supplementary Material online). In addition to the common alr2P1 and alr2P2 duplications, haplotype-specific duplications were also observed. The most prominent example was the presence of alr2P3 on alr2-F (fig. 5A, supplementary fig. 7, Supplementary Material online). Instances of putative transpositions were also observed as homologous nonsynthetic segments found in different haplotypes. Although many examples of duplications and transpositions were identified, a precise history of these rearrangements could not be inferred from the limited number of haplotypes.

Discussion
Diversity of alr2 Alleles
We demonstrated a high degree of alr2 allelic variability—34 of 35 randomly sampled alleles encoded unique gene products and 29 of 35 encoded unique extracellular domains—commensurate with the known allotype diversity. Given the rarity of fusion events observed between field-collected colonies (Grosberg et al. 1996; Mokady and Buss 1996; Nicotra and Buss 2005), the exceptional allelic diversity at alr2 was not unexpected. Of the 301 pairwise comparisons (210 type I, 91 type II) of the unbiased cDNA data set, only one pair of predicted proteins was identical and six pairs (2%) matched over their extracellular domains (fig. 1). This frequency of randomly matching alleles is within the range of 1.8–4.4% fusibility between field-collected colonies (Nicotra and Buss 2005). As nearly all the predicted gene products were unique, the alleles were observed at near-equivalent frequencies. Taken together, a population of many, equally rare alleles is a hallmark of negative frequency–dependent selection (Wright 1968).

Frequency-dependent selection may act to restrict fusibility to close kin that still retain highly similar alr2 alleles (Buss and Green 1985; Grosberg 1988). Colonial organisms like Hydractinia maintain multipotent stem cells but do not segregate their germ lines during embryogenesis (Buss 1982, 1983). When colonies fuse, their germ line becomes susceptible to invading stem cells. Germ line parasitism has been documented in organisms as diverse as prokaryotic myxobacteria (Velicer et al. 2000), eukaryotic cellular slime molds (Buss 1982; Ennis et al. 2000; Strassmann et al. 2000), and the protochordate Botryllus (Stoner and Weissman 1996). If one colony parasitizes the germ line of an unrelated individual, the victim would fail to reproduce and thus its fitness would effectively be zero. Avoiding this risk is thought to account for the predominance of rejection between randomly paired colonies. In the case of parasitism by close kin, however, the fitness cost would be mitigated by a high degree of relatedness or shared genetic background (Buss and Green 1985).

Can the high levels of alr2 heterozygosity be explained by the same selective forces? A diploid genotype derived from a pool of highly diverse alleles is more likely to be heterozygous than if the individual’s genotype were derived from a pool of common alleles. Indeed, all the field isolates were shown to be heterozygous at alr2, except for colony LH07:043 that yielded only a single allele. A substantial effort to clone and sequence a second allele produced a single contig consisting of 41 reads. This colony may either be homozygous, hemizygous, or its other allele is so divergent as to fail to amplify with any of the myriad primers tested. In any case, heterozygosity is the norm. Heterozygote
advantage, or heterosis, at alr2 has not been formally ad-
dressed. Interestingly, fewer than the expected number of
ARC-f homozygous offspring were recovered in a large
backcross including near-congenic laboratory lines (Powell
et al. 2007). This result could be explained by inbreeding
depression, although we cannot exclude the possibility that
alr heterozygosity confers some fitness advantage. De Tom-
aso and Weissman (2004) quantified a disproportionate
number of fuhc (allorecognition locus) heterozygotes in
B. schlosseri and concluded that heterosis (overdominant
selection) maintained fuhc diversity. Heterosis has been
proposed as a driver of MHC heterozygosity as well (Penn
et al. 2002). Whether or not heterozygosity at alr2 confers
a fitness advantage, the extraordinarily large number of
unique alleles suggests that negative frequency–dependent
is a suitable explanation for the observed allele distribution.
If natural selection acts to maintain variation, the question
remains how the variation arises.

Generation of Sequence Diversity
Our analysis of field-derived alr2 alleles has shown that
both sequence and structural variation contribute to
alr2 diversity. Sequence variation is concentrated in the
V-set-like domain I, as previously reported (Nicotra et al.
2009). We confirmed that selection acts to increase alr2
diversity (see fig. 3 in this study and Nicotra et al. 2009),
but do the mutations on which selection acts arise by
any nonneutral mechanisms? If mutations appeared at random
throughout the gene, one would expect the rate of silent
polymorphisms, or π(s), to be roughly the same in
each domain, corresponding to a rate similar to that of
the gene as a whole. Instead, we observed π(s) elevated
1.8- to 2.6-fold in the extracellular region, especially domain
I (fig. 4). This finding indicates that both synonymous and
nonsynonymous mutations arise more frequently in the
hypervariable region.

When the first 1.4–1.8 kb of gDNA sequence from 26
alr2 alleles was analyzed, variation increased markedly in
intron 1 around 150 bp from the start of transcription,
peaked and fell in exon 2, then continued to rise through-
out intron 2 (supplementary fig. 2, Supplementary Material
online). Much of the variation in this large intron could be
attributed to structural differences (i.e., indels, repeats, and
rearrangements), as well as the accumulation of noncoding
mutations in the absence of selection. Nevertheless, the
pattern of variation over the CDSs is reminiscent of that
caused by transcription-linked SHM in vertebrate V-D-J
gen (Steele et al. 1992; Rada and Milstein 2001; Odegard
and Schatz 2006). In the case of SHM, genomic mutations
begin after a “lag” of ~150 bp from transcription start, ac-
cumulate steeply marking the 5’ boundary and decay with
increasing distance from transcription start along a 5’ to 3’
gradient (Steele et al. 1992; Rada et al. 1994; Rada and
Milstein 2001; Odegard and Schatz 2006). A couple of ob-
vious differences exist between SHM and the case of alr2.
First, we did not detect a 3’ decay in variation. From the
amino acid H’ (fig. 3) and cDNA π(s) (fig. 4) data, however,
we know that variation declines in the coding regions after
exon 2. Thus, even if a pattern of decay exists, we may
simply have failed to sample far enough into the gDNA
sequence to detect it. Second, SHM is by definition
a somatic phenomenon. A priori we would not expect
a mutational mechanism to create somatic variants in
Hydractinia as somatic diversity within a colony could lead

Fig. 6. Map of insertions/deletions among alr2 haplotypes. Contigs were compared by MultiPipMaker, with each contig alternatively designated as the reference sequence. In each case, the reference is represented by arrows indicating the location of genes. Conserved regions are shown as pink boxes, less-conserved regions as green boxes, and indels as white boxes (gaps), relative to the reference sequence. Gene synteny is indicated by dotted lines. Putative mobile elements are labeled as in figure 5b.
to auto-intolerance. In fact, no more than two alleles were recovered from any colony (supplementary table 1, Supplementary Material online), consistent with the expectation that variants do not arise in the soma. Thus, if variation in alr2 results from a hypermutagenic mechanism, it would be restricted to the stem cells or occur during meiosis.

Structural Variation at the alr2 Locus
The discovery of two distinct structural types of alr2 alleles revealed another source of variation in addition to sequence polymorphism. Close inspection of specific protein sequences suggested that varying exon structure by recombination may actively contribute to alr2 diversity (fig. 2, supplementary fig. 1, Supplementary Material online). Two sets of alleles possessed nearly matching domains I but varied in structure type. Given the hypervariability of domain I, one would only expect to see highly similar sequences in closely related alleles. Additional evidence of recombination was found in four alleles that possessed type I structural features but type II-specific polymorphisms in domains II and III, encoded by exons 3 and 4. The observation that several type II alleles contained a short amino acid tract found in all sequences of type I further hints at recombination. Recombination could occur as crossing over and/or gene conversion and could take place between alr2 alleles and/or its pseudogenes.

Origin of the alr2 Subcomplex
Analyzing the alr2-F haplotype, Nicotra et al. (2009) suggested that the alr2-related pseudogenes arose from partial tandem duplications of alr2. In the current study, the duplicated genes were found to be structurally related to type II alleles. In addition to the genic similarities, haplotype comparisons further indicated the extent of the duplicated intergenic sequences (supplementary fig. 7, Supplementary Material online). On LH06:003B, alr2P1 resided in an interval with substantial colinearity to >14 kb of intergenic sequence adjacent to the 5′ end of alr2. Similarity between the alr2P2 interval and other alr2-related sequences, meanwhile, spanned only the 5′ promoter to ~3 kb of 3′ flanking sequence. These data suggest that the duplication events were independent and unequal in size. Furthermore, the alr2P2 gene showed greater colinearity with alr2P1 than it did with alr2, suggesting that alr2P1 may have been the template for a subsequent duplication event to generate alr2P2. A third partial duplication of alr2-related sequence (alr2P3), spanning the promoter region into exon 2, was identified on the alr2-F haplotype. Although this sequence was not predicted to encode a functional gene, and no expression was observed in RT-PCR experiments (data not shown), the very presence of alr2P3 indicates that higher order duplications may continue to generate structural variation between alr2 haplotypes.

Structural Heterogeneity
The most obvious variation between haplotypes was in the size of the genomic interval, which ranged from 94 to 118 kb. As this result suggests, indels both large and small were distributed across the haplotypes (figs 5D and 6). These were most prominently concentrated in the intergenic interval between atp synthase and alr2P2/P3, and within the alr2 gene itself. The same intergenic region was also found to contain putative transpositions, that is, sequences that were similar to those found in nonsyntenic locations on other haplotypes. With only three independent genomic contigs, assigning polarity to these changes was not possible. Still, these data suggest that insertions–deletions, duplications–deficiencies, and transpositions all may contribute to structural variation among alr2 haplotypes.

In some well-documented cases, size and structural variation of syntenic regions of gDNA have been attributed to the presence/absence of repeat DNA, as well as to the mobility of some of these elements (Bruggmann et al. 2006; Hawkins et al. 2006). A region of the maize chromosome 9 long arm, for example, was 40% larger than the syntenic interval from chromosome 3 in rice. This difference (7.8 vs. 4.9 mb) was attributed to differential contraction in rice and/or differential expansion in maize from the insertion of transposable elements. Of course, metazoan genomes, including that of the cnidarians Hydra magnipapillata and Nematostella vectensis, are also composed of substantial quantities of transposable elements and repetitive DNA (Putnam et al. 2007; Naamati et al. 2009; Chapman et al. 2010).

Not unexpectedly, the alr2 haplotypes were found to contain putative mobile elements with similarity to reverse transcriptase, copia-type retrotransposon, and tigger transposable element–derived protein, as well as repetitive DNA (figs 5 and 6, supplementary fig. 5, Supplementary Material online). The mobile elements often corresponded to the position of indels. In some cases, elements with highly similar sequences, such as (c) on the A and F haplotypes, were found in nonsyntenic positions flanking the alr2-related genes. These may play a role in further varying the organization of the alr2 subcomplex. Among the repetitive DNA tracts, many belonged to families common to all three haplotypes, whereas other repeats were haplotype specific. Many of the repeat families were found within alr2 and its copies, which, as duplications, are inherently repetitive. Not all the alr2-related repeats were simply duplications, however. Intergenic sequences were identified with strong similarity to segments of intron 2, intron 3, and exon 4. Intron 2 and intron 3 repeats were also found on bacterial artificial chromosome sequences around alr1 and among unassembled reads of another H. symbiolongicarpus genome (data not shown). This suggests that the alr2 region contains repeat families that are more generally distributed throughout the genome.

Repetitive DNA may not be the only source of haplotype diversity. The possibility exists that alr2-related duplications represent templates for recombination-based variation. Several forms of homology-dependent recombination are known to diversify recognition molecules in other systems. Many of the MHC loci display vast allelic diversity generated by gene conversion (Holmes and Parham 1985; Gorski and Mach 1986; Kuhner et al. 1991; von Salome and Kukkonen 2008). Gene conversion has also
been shown to be largely responsible for the diversity of the avian (i.e., chicken) Ig and MHC repertoire (Bezzubova and Buerstedde 1994; Hosomichi et al. 2008). Meiotic recombination, including unequal crossing over, has been implicated in generating diverse lineages of KIRs in primate natural killer cells (Norman et al. 2009). Alternatively, a copy choice mechanism seems to govern the piecwise assembly of variable lymphocyte receptors in the jawless fish (Nagawa et al. 2007).

Pseudogenes have been shown, empirically and theoretically, to contribute to sequence variability of related functional genes in various systems outside of Hydractinia. Gene conversion events with pseudogene donors have been implicated in various diseases caused by many human loci (Bischof et al. 2006). Gene conversion is not always deleterious, however. Pseudogene CDSs can be maintained intact, or even diversified by selection, if their role in gene conversion confers a fitness advantage (Takuno et al. 2008). The identification of two putative recombination tracts in haplotype A alr2P genes, although not conclusive in origin, indicated the occurrence of some form of recombination between alr2-related sequences. The speculation that the alr2P genes could be sequence donors or templates for intergenic recombination or gene conversion requires further exploration. Deeper haplotype sampling could reveal more instances of recombination and may allow for more reliable designation of donor and recipient sequence groups. A potential role for the alr2P genes in allodiversity is an attractive conjecture that would be consistent with mechanisms known to vary recognition specificity in other systems.

Conclusions

Separated by hundreds of millions of years of evolution, Hydractinia and higher animals both exhibit a striking consistency in their reliance on allelic diversity to determine histocompatibility. The Hydractinia allodeterminant alr2 displays extremely high natural genetic variability, commensurate with known allotype diversity. Variation is achieved by both sequence and structural polymorphism, and novel alleles are likely favored by negative frequency-dependent selection. Genomic analyses indicate a rich history of duplications, insertions and deletions, and rearrangements in the alr2 region that may also contribute to diversity at the level of haplotype. We propose that the mechanisms generating this variation may include interallelic recombination, transcription-linked hypermutation, and/or gene conversion.

Supplementary Material

Supplementary tables 1–3, figs 1–7, Alignment File, and Results and Discussion are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank K. Altland, C. Glastris, S. Kirschner, C. Anderson, and E. Buss for technical assistance. S. Rosa, M. Nicotra, A. Gloria-Soria, A. Powell, F. Lopez, J. Townsend, and D. Schatz provided useful discussion. This paper has been greatly improved thanks to the comments of two anonymous reviewers and the associate editor. The work described herein was supported by the National Science Foundation (IOS-0818295 to L.W.B, S.L.D, and M.A.M.) and National Institutes of Health (R56-0006579/11692-1 to F.G.L., L.W.B, and S.L.D.). R.D.R. was supported, in part, by the National Science Foundation predoctoral Graduate Research Fellowship. Computational analyses were performed on the Yale University Biomedical High Performance Computing Center, which is supported by the National Institutes of Health (RR19895).

References


Downloaded from https://academic.oup.com/mbe/article-abstract/28/2/933/1218983 by guest on 08 November 2018


