Evolutionary Genetics of the Hydroid Allodeterminant alr2

Andrea Gloria-Soria,1 Maria A. Moreno,2 Philip O. Yund,4,5 Fadi G. Lakkis,4,5 Stephen L. Dellaporta,2 and Leo W. Buss*1

1Department of Ecology and Evolutionary Biology, Yale University
2Department of Molecular, Cellular and Developmental Biology, Yale University
3Marine Science Education and Research Center, University of New England
4Department of Surgery, Thomas E. Starzl Transplantation Institute, University of Pittsburgh School of Medicine
5Department of Immunology, Thomas E. Starzl Transplantation Institute, University of Pittsburgh School of Medicine

*Corresponding author: E-mail: leo.buss@yale.edu.

Abstract

We surveyed genetic variation in alr2, an allodeterminant of the colonial hydroid Hydractinia symbiolongicarpus. We generated cDNA from a sample of 239 Hydractinia colonies collected at Lighthouse Point, Connecticut, and identified 473 alr2 alleles, 198 of which were unique. Rarefaction analysis suggested that the sample was near saturation. Most alleles were rare, with 86% occurring at frequencies of 1% or less. Alleles were highly variable, diverging on average by 18% of the amino acids in a predicted extracellular domain of the molecule. Analysis of 152 full-length alleles confirmed the existence of two structural types, defined by exons 4–8 of the gene. Several residues of the predicted immunoglobulin superfamily-like domains display signatures of positive selection. We also identified 77 unique alr2 pseudogene sequences from 85 colonies. Twenty-seven of these sequences matched expressed alr2 sequences from other colonies. This observation is consistent with pseudogenes contributing to alr2 diversification through sequence donation. A more limited collection of animals was made from a distant, relict population of H. symbiolongicarpus. Sixty percent of the unique sequences identified in this sample were found to match sequences from the Lighthouse Point population. The large number of alr2 alleles, their degree of divergence, the predominance of rare alleles in the population, their persistence over broad spatial and temporal scales, and the signatures of positive selection in multiple residues of the putative recognition domain paint a consistent picture of negative-frequency-dependent selection operating in this system. The genetic diversity observed at alr2 is comparable to that of the most highly polymorphic genetic systems known to date.

Key words: polymorphism, allorecognition, Hydractinia, sequence donor, population genetics, balancing selection.

Introduction

Allorecognition is the capacity to discriminate between conspecífics based on cell–cell contact. Allorecognition systems regulate mating interactions in fungi (Kämper et al. 1995; Fraser and Heitman 2003), social recognition in amoebae (Benabentos et al. 2009; Hirose et al. 2011), and inbreeding in plants (Charlesworth 2010). In vertebrates, the major histocompatibility complex (MHC) is a determinant of allograft acceptance (Klein 1986). Among colonial marine invertebrates, allorecognition restricts colony fusion to self or close kin, thereby reducing the costs of germline parasitism (Buss 1982).

Loci that encode recognition molecules are characterized by extreme allelic diversity. Hundreds of alleles have been identified at loci from the human MHC (Robinson et al. 2000, 2011), whereas the plant self-incompatibility (SI) loci can carry as many as 50 alleles within a population (Charlesworth 2010). Parasites such as Trypanozoma brucei or Plasmodium falciparum can host tens to hundreds of alleles for the cell surface proteins commonly targeted by the human immune system (Pays et al. 1983; Kast et al. 1994; Hoffmann et al. 2001; Aubouyet al. 2003; Taylor and Rudenko 2006; Ghanchi et al. 2010).

Extraordinary polymorphism can be accounted for if alleles are generated frequently, if balancing selection preserves alleles in the population over time, or both (Wright 1939; Fisher 1958; Lewontin et al. 1978). Novel diversifying mechanisms are known for several highly variable systems. The primary antibody repertoire in humans and mice is produced by combinatorial rearrangement and imprecise joining of gene fragments, followed by the addition of nontemplated nucleotide insertions, also known as V(D)J recombination (Alt and Baltimore 1982; Charbonnier et al. 1980; Alt and Baltimore 1982; Jenne et al. 2003, 2006). These antibodies are further diversified upon antigen challenge by the stepwise introduction of point mutations through somatic hypermutation (SHM), as well as by gene conversion and class switch recombination. In birds, cattle, swine, and rabbits, where the variability of the V and J segments is limited, diversity is produced by gene conversion events that involve the unidirectional transfer of gene segments from pseudogenes (Reynaud et al. 1987; Becker and
Hydroid-covered hermit crab shells were collected from the intertidal zone at Lighthouse Point (41° 14′ N, 72° 54′ W) during the summer and fall of 2009. The crabs were removed from their shells by bubbling CO₂ into the aquarium. A second collection of hermit crabs was made in the shallow subtidal of Wolf Neck, Maine (43° 49′ N, 70° 04′ W) during the fall of 2010. Two sibling species, *H. symbiolongicarpus* and *H. polyclina*, are found within the Gulf of Maine (Buss and Yund 1989). Colonies of *H. symbiolongicarpus* were identified by 16S ribosomal DNA sequence, using primers P1982 and P1983 (**Table 1**) and Phusion Hot-Start polymerase (Finzymes), under the following conditions: 98°C × 30 s, 30X (98°C × 10 s; 62°C × 30 s; and 72°C × 45 s), 72°C × 4′.

**Materials and Methods**

**Sample Collection**

Total nucleic acids were extracted using a modified urea extraction protocol (Chen and Dellaporta 1994). RNA was isolated from the nucleic acid preparations by treating with recombinant RNase-free DNase I (Roche), followed by heat inactivation at 75°C and the addition of EDTA to a final concentration of 0.1 M. Approximately 300 ng of total RNA was used as a template for first-strand *alr2*-specific cDNA synthesis using the SMARTScribe reverse transcriptase (Clontech) and the P3981 primer (**Table 1**), following manufacturer instructions. Primer P3981 was designed to bind at *alr2* exon 9 to ensure that only cDNA from full-length *alr2* messages was synthesized, thus eliminating both incomplete messages and pseudogenes from the cDNA pool.

**Amplification and Barcoding of *alr2* Alleles**

Primers used to generate *alr2* sequences are given in **Table 1**. Full-length coding sequences (CDSs) were amplified from first-strand cDNA in a single polymerase chain reaction (PCR) using primers P3980 and P3967 with Phusion Hot-Start High-Fidelity DNA polymerase (Finzymes), according to manufacturer instructions. The thermocycler conditions were 98°C × 1′, 5X (98°C × 10 s, 68°C × 30 s, 72°C × 1′15″), 5X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″), 25X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″), 7.5X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″), 6X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″), 5X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″), 4X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″), and 3X (98°C × 10 s, 67°C × 30 s, 72°C × 1′15″).
A unique combination of four nucleotides, or barcode, was included to facilitate downstream cloning (Aslanidis and de Jong 1990). A PCR amplicon and links each allele to its original colony. The barcode is integrated to the final ligation-independent cloning (LIC) sequences to serve as a colony identifier. The barcode is translated using Geneious v. 5.3 (Drummond et al. 2011) and custom R script and then curated, assembled, aligned, and translated using MUSCLE (Edgar 2004). The consensus sequence was used in cases where alr2 sequences from the same individuals differed by five or fewer base pairs. Sequences bearing barcodes that did not correspond to their pool were discarded. Analysis included 39 additional full-length sequences from Rosengarten et al. (2011). All sequences are provided as an alignment file in the supplementary materials, Supplementary Material online, and in GenBank (JX048709–JX049092). Unique allele sequences were identified using DnaSP v. 5.10.01 (Librado and Rozas 2009). Pairwise comparisons were performed using PAUP v. 4.0 (Swofford 2003) and excluded gaps. Sequence diversity of alr2 alleles was quantified by generating a pairwise distance matrix. The resulting nucleotide distances were normalized to sequence length and used to generate an identity heat map (identity = 100 × (1 − normalized distance)). Sampling effort was assayed by simulating a rarefaction curve with the rarefaction calculator from Brzustowski (1998). Per-site variability was estimated with the Shannon Entropy-One application from the HIV Sequence Database (2005). Bayesian analyses were conducted with MrBayes 3.2.1 (Huelsenbeck and Ronquist 2003) using default priors, conducted to with MrBayes 3.2.1 (Huelsenbeck and Ronquist 2003) using default priors, with four incrementally heated Markov chains, and two concurrent runs of 2 × 10^6 generations sampled every 1 × 10^4 generations. The best fitting model for each region was determined using Akaike information criterion as implemented by Modeltest 3.7 (Posada and Crandall 1998). The models used were exons 1, 4, 6, and 9 (Hasegawa-Kishino-Yano [HKY] + I + G); exon 2, V sets, and full-length sequences (general time reversible [GTR] + I + G); exon 3 (HKY + I + G); exon 5 (GTR + G); exon 7 (HKG); and exon 8 (GTR). Plots of log-likelihood scores versus generation time were used to assess stabilization and convergence of the Markov chains. Trees generated before stabilization were discarded. In all cases, stabilization was achieved by generation 200,000 (20 sampled trees).

### Table 1. Nucleic Acid Primers.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Type</th>
<th>Sequence</th>
<th>Direction</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1982</td>
<td>Species typing</td>
<td>TCGACTGTTACCAAAACATA</td>
<td>fwd</td>
<td>16S rDNA</td>
</tr>
<tr>
<td>P3967</td>
<td>Species typing</td>
<td>ACGGAATGAATCAGGATGTAAG</td>
<td>rvs</td>
<td>16S rDNA</td>
</tr>
<tr>
<td>P3980</td>
<td>PCR</td>
<td>tacctcaccactACCAATATGAGCTGAAAC</td>
<td>fwd</td>
<td>Exon 1</td>
</tr>
<tr>
<td>P3981</td>
<td>RT-PCR</td>
<td>TACCGAGGCTGCAACACTTGTG</td>
<td>rvs</td>
<td>Exon 9</td>
</tr>
<tr>
<td>P4028</td>
<td>Sequencing</td>
<td>GCTCTGTAGGTAATTCAGACCA</td>
<td>fwd</td>
<td>Exon 4</td>
</tr>
<tr>
<td>P4032</td>
<td>Sequencing</td>
<td>ATTATTATGGKGGTCCAA</td>
<td>rvs</td>
<td>Vector</td>
</tr>
<tr>
<td>P4088</td>
<td>sequencing</td>
<td>AGTGGAAACCTCTACGTGCC</td>
<td>fwd</td>
<td>Vector</td>
</tr>
<tr>
<td>P4089</td>
<td>Sequencing</td>
<td>CACCCGCGGAGGCTAGGCC</td>
<td>rvs</td>
<td>Vector</td>
</tr>
<tr>
<td>P4494</td>
<td>PCR</td>
<td>tatccaccttactgCATAAATAATYARMTTYGTAGTGCTATT</td>
<td>rvs</td>
<td>Exon 2</td>
</tr>
</tbody>
</table>

**Note.**—rDNA, ribosomal DNA.

**LIC sites are in lower case.**

**Primer direction: fwd, forward and rvs, reverse.**

**TSL, trans-splice leader sequence.**

**Exon boundary.**

**Vector pYU2418.**

Pooling, Cloning, and Sequencing of alr2 Alleles

The alr2 amplicons were normalized and pooled in groups of eight. Each pool was gel purified following the QIAquick gel-extraction kit protocol (Qiagen). Purified pools were then cloned in chemically competent Top10 cells (Invitrogen) by LIC-cloning (Aslanidis and de Jong 1990) into custom LIC-adapted plasmid pYU2418. Transformants were selected on LB agarose plates with kanamycin (30 μg/ml) and confirmed by PCR. Sequences of the 5′-UTR, exon 1, and hypervariable exon 2 were obtained using vector primer P4088. Full-length alr2 sequences were obtained using primers P4088, P4028, P4032, and P4089. Sanger sequencing was performed at the University of Washington High-Throughput Genomics Unit (Seattle, WA).

This protocol was also applied to H. symbiolongicarpus and H. symbiolongicarpus/C22, 72° × 5′. Both P3980 and P3967 included ligation-independent cloning (LIC) sequences to facilitate downstream cloning (Aslanidis and de Jong 1990). A unique combination of four nucleotides, or barcode, was added to the 3′-end of the LIC sequence within P3980 to serve as a colony identifier. The barcode is integrated to the final PCR amplicon and links each allele to its original colony.

Sequence Analysis

Chromatograms were sorted based on their barcode using a custom R script and then curated, assembled, aligned, and translated using Geneious v. 5.3 (Drummond et al. 2011) and MUSCLE (Edgar 2004). The consensus sequence was used in cases where alr2 sequences from the same individuals differed by five or fewer base pairs. Sequences bearing barcodes that did not correspond to their pool were discarded. Analysis included 39 additional full-length sequences from Rosengarten et al. (2011). All sequences are provided as an alignment file in the supplementary materials, Supplementary Material online, and in GenBank (JX048709–JX049092). Unique allele sequences were identified using DnaSP v. 5.10.01 (Librado and Rozas 2009). Pairwise comparisons were performed using PAUP v. 4.0 (Swofford 2003) and excluded gaps. Sequence diversity of alr2 alleles was quantified by generating a pairwise distance matrix. The resulting nucleotide distances were normalized to sequence length and used to generate an identity heat map (identity = 100 × (1 − normalized distance)). Sampling effort was assayed by simulating a rarefaction curve with the rarefaction calculator from Brzustowski (1998). Per-site variability was estimated with the Shannon Entropy-One application from the HIV Sequence Database (2005). Bayesian analyses were conducted with MrBayes 3.2.1 (Huelsenbeck and Ronquist 2003) using default priors, conducted to with four incrementally heated Markov chains, and two concurrent runs of 2 × 10^6 generations sampled every 1 × 10^4 generations. The best fitting model for each region was determined using Akaike information criterion as implemented by Modeltest 3.7 (Posada and Crandall 1998). The models used were exons 1, 4, 6, and 9 (Hasegawa-Kishino-Yano [HKY] + G); exon 2, V sets, and full-length sequences (general time reversible [GTR] + I + G); exon 3 (HKY + I + G); exon 5 (GTR + G); exon 7 (HKG); and exon 8 (GTR). Plots of log-likelihood scores versus generation time were used to assess stabilization and convergence of the Markov chains. Trees generated before stabilization were discarded. In all cases, stabilization was achieved by generation 200,000 (20 sampled trees).
Prior work had identified structural polymorphisms that define two different classes of alr2 alleles (Rosengarten et al. 2011). Selection tests were performed on each class separately. Recombination breakpoints were estimated using the GARD algorithm from the Data Monkey webserver (Kosakovsky-Pond and Frost 2005; Kosakovsky-Pond et al. 2005; Delport et al. 2010). Sites under selection were identified using the log ratio of nonsynonymous [Pi(s)] to synonymous [Pi(s)] nucleotide diversity per site, as calculated in DnaSP (Librado and Rozas 2009).

Pseudogenes

Two to three pseudogenes that include alr2 exons 1–4 or 5 are physically linked to the alr2 locus in all genomic intervals examined to date (Nicotra et al. 2009; Rosengarten et al. 2011). We generated pseudogene sequences from a sample of 69 Hydractinia colonies for which two alr2 full-length alleles had been identified from cDNA and where we were certain that no additional alleles were present, based on the number of clones sequenced for that individual colony (mean = 17; median = 15). The alr2 region encompassing exons 1 and 2 from these colonies was amplified from genomic DNA. The PCR reaction was performed with Phusion Hot-Start High-Fidelity DNA polymerase (Finnzymes) using primers P3402 and P4494 (table 1) at 98°C × 30s, 30X(98°C × 10s, 62°C × 30s, 72°C × 45s), 72°C × 10s. Those sequences recovered from the genomic amplification but absent from the cDNA pool were considered alr2 pseudogenes. Identification of pseudogenes was achieved by comparing the sequences within each individual colony, followed by the removal of sequences matching the expressed alleles (≤5 bp mismatches). Thirty-two pseudogenes identified from Rosengarten (2010) were also included in this analysis. All pseudogenes were compared to the alr2 5'-unique sequences using pairwise distance analysis in PAUP v. 4.0 (Swofford 2003).

A summary of all sequences presented here is provided in supplementary table S1, Supplementary Material online.

Results

Natural alr2 Diversity

To study the amount and distribution of natural genetic variation at the H. symbiolongicarpus alr2 locus, we analyzed 239 colonies collected at Lighthouse Point (LH09). The region of alr2 that encompasses exons 1 and 2 was sequenced. This interval included the putative signal peptide, encoded by exon 1, and the hypervariable domain of alr2, encoded by exon 2.

Sorting of the alr2 sequences using barcodes resulted in the identification of 473 alr2 alleles, 10-fold more than sequenced in previous studies (Nicotra et al. 2009; Rosengarten et al. 2010). We recovered two alleles from 150 of the colonies (~63%) and one allele from 54 colonies (~23%). In addition to the two alleles expected from a single copy gene in heterozygote state, extra alleles were recovered from 35 (15%) of the colonies. The frequency distribution of the alleles recovered per colony is given in supplementary figure S1, Supplementary Material online.

Next, we estimated the number of unique alr2 alleles represented within the sample. Pairwise comparisons of all 473 alleles identified 198 unique nucleotide sequences, encoding 180 amino acid sequences (supplementary table S2 and alignment files, Supplementary Material online). Unique sequences will be referred to as haplotypes in what follows. Eleven of the haplotypes contained stop codons and were removed from the data set. The large number of alr2 alleles found at Lighthouse Point indicates that alr2 is a highly polymorphic locus.

The sequence diversity of alr2 haplotypes was quantified in a pairwise distance matrix and used to generate the identity heat map displayed in figure 1A. The average nucleotide distance between haplotypes was 54 out of 571 bases, with a maximum pairwise distance of 97 nucleotides (median = 57; fig.1B). The predicted protein sequences diverged by an average of 33 amino acids, out of 184 possible, with the most divergent sequences differing by 61 residues (median = 34; fig. 1C). These results indicate that alr2 is highly variable.

Next, we examined the distribution of allele frequencies in the LH09 population. This step was possible thanks to the barcoding strategy, which identified the alleles originating from a single colony and allowed the detection of independent colonies with shared alleles. The frequency distribution, depicted in figure 2, showed a bias in the allele distribution toward low frequencies. Fifty-four percent of the haplotypes were singletons (frequency f = 0.002). In contrast, haplotypes 12, 19, and 41 were found at frequencies an order of magnitude higher (f = 0.025–0.045; supplementary table S2, Supplementary Material online). Rare alleles were common in LH09, and shared alleles occurred only sporadically.

We assessed how well our LH09 sample captured alr2 diversity using a rarefaction curve simulated in a rarefactor calculator (Brzustowski 1998). The algorithm randomly draws virtual subsamples from the actual data set and plots a curve based on the average number of alleles found per sample size. A steep slope indicates that many alleles are yet to be discovered, whereas a flattening curve suggests that only rare alleles remain to be identified in the population. The rarefaction analysis indicated that the curve begins to flatten at the sample size we analyzed (fig. 3). Therefore, our sampling of allelic diversity was near saturation for this location.

We also recovered 24 alr2 alleles from 17 H. symbiolongicarpus colonies collected in Wolf Neck. Eleven of these alleles were unique within this population, and six of these differed by five base pairs or less from alleles sampled at the Lighthouse Point (supplementary table S2 and alignment file, Supplementary Material online).

Full-Length Sequences

Full-length CDSs were used to further investigate the source and pattern of genetic variation in alr2. The clones of 119 of the haplotypes were sequenced for all nine exons and analyzed together with 39 of the full-length sequences obtained by Rosengarten et al. (2011). Translation of the CDSs...
detected six sequences with premature stop codons, which were excluded from subsequent analyses. Two main subgroups were apparent from the alignment of the 152 remaining CDSs (n = 1,906 characters; alignment file, Supplementary Material online). One group (Type I) codon on average 616 amino acids, and the second group (Type II) encoded on average 606 amino acids. Type I alleles had a mean pairwise identity of 95.6% (n = 90) and Type II of 94.1% (n = 62).
The two allele subtypes were in agreement with the previous Type I and II designations from Rosengarten et al. (2011).

To further investigate the distribution of Type I and II alr2 alleles in the population, we analyzed the full-length sequences from 62 of the newly reported colonies bearing two alleles. Two Type I alleles were present in 18 colonies, two Type II alleles occurred in 13, whereas 31 colonies bore one allele of each type. These frequencies are not significantly different from that expected by random assortment at a single locus ($\chi^2$ test; $P = 0.99$).

To study the two alr2 subtypes in more detail, we generated Bayesian trees from the full-length CDS and from each of the individual exons (fig. 4 and supplementary fig. S3, Supplementary Material online). Two clusters, consistent with the Type I and Type II subdivisions were observed in the full-length tree (fig. 4A). Interestingly, although Type I and Type II clusters were observed in the trees from exons 4 to 9, the distribution of exons 1 and 2 in the tree was not type specific (fig. 4B and supplementary fig. S3, Supplementary Material online). Moreover, in the tree of exon 3, three clusters were identified, one with Type I alleles, a second exclusively containing Type II, and a third intermediate clade that contained both Type I and Type II alleles. The distribution of alleles on the Bayesian trees supported the presence of two alr2 allele types, defined primarily by type-specific exons 4–8, with a partial contribution of exon 3.

We addressed whether the subdivision in alr2 alleles was indicative of two functional classes by analyzing the evolutionary fingerprints of Type I and Type II alleles. The evolutionary fingerprint model assumes that each molecule has a characteristic selection pattern that depends on its function and structure (Delport et al. 2010). Molecules with similar functions are expected to experience similar evolutionary forces and thus will have similar evolutionary fingerprints. Direct comparison of the fingerprints indicated that despite their structural polymorphisms, the evolutionary fingerprints of Type I and Type II alleles were similar. Both of the fingerprints were characterized by the presence of two evolutionary rates of base substitutions along the molecule (supplementary fig. S2, Supplementary Material online), indicating that Type I and Type II alleles do not represent different functional classes.

We then tested for signatures of the diversifying mechanism operating on alr2. Variability at each nucleotide position of the alignment was evaluated by the Shannon entropy index ($H'$). This index represents the uncertainty that a nucleotide will be present at a specific site (Cover and Thomas 2001). Low entropy values are indicative of low diversity, and high entropy values suggest high diversity. The distribution of entropy values for Type I and Type II alleles is displayed in figure 5, and the average entropy per exon is presented in table 2. The two structural types display different patterns of variability. Exon 2 was the most variable exon in both allele types. In Type II, high variation ($H' > 0.1$) extended through exon 4, whereas in Type I, diversity decayed soon after exon 2 (table 2, fig. 5). No significant difference in average diversity was detected between allele types ($t$-test $P = 0.057; df = 1$). This suggests that variability distribution along alr2 is not random and that Type I and Type II alleles differ in their diversity pattern.

To address whether the alr2 allele data set included recombinant sequences, we applied the GARD algorithm to the CDSs alignment (Kosakovsky-Pond et al. 2006). GARD defines nonrecombinant fragments by searching the sequence alignment for segment-specific phylogenies. GARD located three breakpoints at positions 408, 817, and 1332 in Type I alleles ($P < 0.001$) and two breakpoints at nucleotides 415 and 1145 in Type II alleles ($P < 0.001$) (fig. 5). This suggests that at least some of the alr2 alleles are the product of recombination events and supports a differential pattern of nucleotide diversity between the different allele types.

We then identified sites under selection by estimating the log ratio of nonsynonymous to synonymous nucleotide diversity [$\log(Pi(a)/Pi(s))$] per site along the molecule (fig. 5). Type I and Type II alleles were analyzed independently. Values above zero indicate sites under positive selection, whereas values below zero suggest negative selected sites. All positive selected sites were found between exons 1 and 5, with the exception of a site located in exon 9 in Type II.
alleles. The majority of the sites under positive selection localized to exon 2 in both of the allele types (T1: 75% and T2: 43%, fig. 5). The localization of positive selected sites to exon 2 was consistent with previous reports (Nicotra et al. 2009; Rosengarten et al. 2011) and with exon 2 being a determinant in the allorecognition event.

Pseudogenes and Sequence Donation
We assessed the possibility that alr2 pseudogenes could act as donor sequences and promote allelic diversification by identifying pseudogene sequences that matched alr2 alleles. The pseudogenes are physically linked to the alr2 locus and include alr2 exons 1–4 or 5 (Nicotra et al. 2009; Rosengarten et al. 2011), thus the alr2 variable extracellular domain. We identified a total of 100 pseudogene sequences from a sample of animals for which only two expressed alleles were known to occur. These included 69 colonies from our LH09 collection and 16 colonies from a previous study (Rosengarten 2010). We consider pseudogenes to be the alr2-like sequences amplified from genomic DNA that were not present in the cDNA pool. Seventy-seven unique pseudogene haplotypes were identified. After intron sequences were removed, the pseudogenes were compared against the alr2 haplotype database, and their pairwise distances were estimated. Sequences differing by five or less single-nucleotide polymorphisms were considered matches. Eleven of these sequences carried one or more stop codons within the two exons analyzed. Results indicated that 27 pseudogene haplotypes matched at least one expressed alr2 allele from a different colony (fig. 6 and supplementary table S2, Supplementary Material online). It is important to note that the criterion to identify pseudogenes is based on their difference with the alr2 alleles expressed in the same colony. Therefore, this approach is not able to detect pseudogenes that more closely match expressed alleles within a colony. The fact that 35% of all pseudogenes (n = 100) were found to match expressed sequences and that pseudogenes and alr2 expressed alleles cluster on the same branches of the Bayesian consensus tree (fig. 6) suggests that sequence donation is taking place between the pseudogenes and alr2.

Discussion
Sampling and Heterozygosity
We have characterized alr2 allelic variation in a population of H. symbiolongicarpus at Lighthouse Point. Rarefaction analysis indicates that sampling of alr2 allelic diversity in this population was near saturation. We recovered two or more alleles from 77% of the colonies surveyed, indicating that high levels of heterozygosity are present. If we assume that the population is in Hardy–Weinberg equilibrium, we would expect very few homozygous animals (~1%). However, only one allele was identified from 54 animals. Certain alleles might be preferentially amplified by the PCR and therefore over-represented in the sequence pool. Biases in the cloning of alleles could also explain the missing alleles. Alternatively, very divergent alr2 alleles may exist in the population that escaped detection by our methods. Our amplification strategy relies on a conserved trans-splice leader (TSL) sequence to
recover the \(alr2\) alleles. Those alleles lacking a TSL or with variation at this sequence may have evaded detection. Finally, the failure to amplify one allele might arise if alleles were differentially expressed. Differential expression of the \(alr2\) alleles is unlikely, as fusibility assays are consistent with a codominant expression (Hauenschild 1954, 1956; Mokady and Buss 1996; Cadavid et al. 2004; Powell et al. 2007).

Unexpectedly, 15% of the \(Hydractinia\) colonies surveyed in LH09 yielded more than two \(alr2\) alleles. An excess of alleles could be explained if some colonies were chimeric (i.e., a product of fusion between two or more colonies that shared an \(alr2\) allele). Fusibility in a nearby population was estimated to occur at a frequency of 3% (Nicotra and Buss 2005), a much lower rate than the frequency at which we observed excess alleles. Tests to detect contamination during sample processing or plate cross-contamination before barcoding were negative. Another possibility is that \(alr2\) displays copy number variation in some wild isolates. Southern analysis of two haplotypes (Nicotra et al. 2009) and complete genomic sequences of the \(alr2\) chromosomal interval for three additional haplotypes (Rosengarten et al. 2011), however, detected no variation in copy number. Moreover, in colonies for which only two alleles were detected, the relative frequencies of Type I and Type II alleles do not differ from the expectation of single locus random assortment. Although we inspected shells to identify single colonies, a final possibility is that multiple colonies were present on the original gastropod shell and went undetected upon initial sampling.

Levels of Variation

The diversity at the \(alr2\) locus is comparable to that of highly polymorphic molecules used by other self-recognition systems. By comparison, plant SI loci bear a maximum of 50 alleles in each species (Charlesworth 2010). The HLA-A allele database, a product of a long-term multilaboratory collaboration, recognizes approximately 1,200 alleles at the human HLA-A locus worldwide (Robinson et al. 2011). Here, we describe 198 \(alr2\) alleles recovered in one collection, from a single \(Hydractinia\) population. Our data suggest that \(alr2\) may be one of the most polymorphic loci known to date.

Maintenance of Variation

The levels of polymorphism observed in \(alr2\) can only be maintained through balancing selection (Wright 1939; Fisher 1958; Kimura and Crow 1964). Balancing selection can maintain variation through heterozygote advantage, habitat heterogeneity, or NFDS. Even if heterozygote colonies at \(alr2\) could benefit from an increase in fusion probability, heterosis alone can only maintain a few alleles in a population (Lewontin et al. 1978). Habitat structure is also unlikely to contribute to the high polymorphism at \(alr2\). LH09 colonies were collected from within 200 m\(^2\) of the intertidal zone, where hydroid-covered hermit crabs move freely and thus no structure would be expected. On the other hand, it has been proposed that \(Hydractinia\) allore cognition acts to prevent somatic parasitism by limiting fusion to close kin (Buss 1982). Movement of interstitial (germline) cells within a chimeric colony can result in one genotype monopolizing the reproductive resources from the other, thus lowering its fitness. Under these circumstances, rare alleles would be favored, as they would decrease the chances of fusion with

Fig. 5. \(Alr2\) diversity plot displaying nucleotide variability along the \(alr2\) CDS. Gray bars represent site variability estimated by the Shannon entropy index (\(H\)). Black bars are the log ratio of nonsynonymous \([Pi(a)]\) to synonymous \([Pi(s)]\) nucleotide diversity at the site. Positive values suggest positive selection, and negative values suggest negative selection. Circles indicate regions of recombination breakpoints as estimated by GARD (\(P < 0.001\)), for Type I (A) and Type II (B) alleles. The bar at the bottom of each graph indicates the position of the exons.

Table 2. \(Alr2\) Exon Diversity (\(H\)).

<table>
<thead>
<tr>
<th>Exon</th>
<th>Type I</th>
<th>Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0784</td>
<td>0.1210</td>
</tr>
<tr>
<td>2</td>
<td>0.2337</td>
<td>0.2206</td>
</tr>
<tr>
<td>3</td>
<td>0.0739</td>
<td>0.1062</td>
</tr>
<tr>
<td>4</td>
<td>0.0387</td>
<td>0.1090</td>
</tr>
<tr>
<td>5</td>
<td>0.0360</td>
<td>0.0916</td>
</tr>
<tr>
<td>6</td>
<td>0.0275</td>
<td>0.0206</td>
</tr>
<tr>
<td>7</td>
<td>0.0055</td>
<td>0.0000</td>
</tr>
<tr>
<td>8</td>
<td>0.0213</td>
<td>0.0144</td>
</tr>
<tr>
<td>9</td>
<td>0.0132</td>
<td>0.0191</td>
</tr>
</tbody>
</table>
an unrelated genotype (Buss 1982; Buss and Green 1985). This suggests that balancing selection might be maintaining the alr2 polymorphism though NFDS.

Multiple lines of evidence suggest that NFDS is acting on alr2. NFDS favors the persistence of rare alleles at low relative frequencies in the population (Wright 1939; Fisher 1958; Kimura and Crow 1964). More than half of the unique alr2 alleles recovered were singletons. Prevalence of rare alleles in the population would result in a large number of alleles and consequently lead to high levels of heterozygosity. We found 198 alr2 alleles at the LH09 population, and a high percentage of the colonies were found as heterozygotes. Alleles maintained by NFDS tend to be very divergent (Takahata and Nei 1990; Charlesworth 2006). Selection for diversity at alr2 is evident from the signatures of positive selection at multiple residues in exon 2, the domain hypothesized to account for allore cognition events. The alr2 alleles from LH09 differ, on average, by 18% of the amino acids in their variable region. Finally, systems under NFDS display alleles with broad geographic distribution and temporal persistence (Klein et al. 1998; Muirhead et al. 2002; Charlesworth 2006). The H. symbiolongicarpus population in Wolf Neck is separated by approximately 600 miles from Lighthouse Point which is an appreciable distance for an animal with a nonfeeding hydroid larva that can crawl but cannot swim. Moreover, the H. symbiolongicarpus population at Wolf Neck is a relict that has been genetically isolated from southern populations for at least 3,000 years (Bousfield and Thomas 1975; Folino and Yund 1998). Our observation that 55% of the Wolf Neck alleles matched alleles from Lighthouse Point is further evidence that NFDS is acting at this locus.

**Generation of Genetic Diversity**

A hallmark of highly variable genetic systems is not only the action of NFDS but also donation of sequences to expressed alleles from a library of unexpressed sequences. The archetypal example of a regulated process is the combinatorial recombination (VDJ like) and class-switch recombination events that generate immunoglobulin diversity in the vertebrate immune system (Janeway et al. 2001). Trypanosome coat proteins are another example. Here, only 1 out of approximately 20 telomeric coat protein (VSG) loci is expressed at a time. Variation in coat protein arises from switching expression to a different VSG telomeric locus or through DNA rearrangements between the expressed gene and silent subtelomeric VSG gene arrays, a process that is itself mutagenic (Barry and McCulloch 2001). Hyphal fusion in filamentous fungi (e.g., Podospora, Neurospora, and Aspergillus) is controlled by multiple het genes, several of which contain highly polymorphic WD repeats (Glass et al. 2000; Saupe 2000; Glass and Kaneko 2003). New allotypes are generated.
by expansion, contraction, and gene conversion within the gene family, as well as by a fungal-specific mutagenesis process known as repeat-induced point-mutation (Paolletti et al. 2007). Gene conversion is the common mode of diversity generation in many of these systems. Gene conversion is also known to diversify the immunoglobulins of birds (Bezzuobova and Buerstedde 1994; Hosomichi et al. 2008) and tetrapods (Holmes and Parham 1985; Gorski and Mach 1986; Kuhner et al. 1991). In Hydractinia, sequence donation has been suggested to contribute to the generation of diversity in alr1, an allodeterminant linked to alr2 in the ARC. Alr1 is found within a gene complex of at least 11 structurally similar IgSF-like genes (Rosa et al. 2010). Sequence analysis of the alr1 complex provided evidence that alleles are shared among these loci and that in at least some cases, gene conversion tracts have been detected (Rosa et al. 2010). A similar process may be involved in the diversification of alr2. The alr2 genomic interval contains several alr2-like sequences that are not part of the expressed gene pool. These pseudogenes include alr2 exons 1–4 or 5 and, therefore, the hypervariable extracellular domain (Rosengarten et al. 2011). Bayesian trees built from individual exons imply that exons 1, 2, and 3 have a different phylogenetic history than the rest of the sequence. This pattern is consistent with a process whereby sequences are shared between pseudogenes and expressed alleles. This suggestion is confirmed by the observation that 35% of all pseudogene sequences match expressed allele sequences. Together, these results demonstrate that alr2 variability is enhanced by pseudogene sequence donation.

**Supplementary Material**

Supplementary tables S1 and S2, supplementary figures S1–S3, and alignment file are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors thank Evan Buss and Stephanie Galarza for technical assistance in the field and laboratory, respectively. R.D. Rosengarten and M.L. Nicotra provided comments. A.G.-S. was supported by a Yale Institute for Biospheric Studies—Donnelley Postdoctoral Fellowship. This work was supported by the National Science Foundation (IOS-0818295 and OCE-09-61157).

**References**


