doi:10.1111/j.1558-5646.2012.01787.x

EVIDENCE FOR SELECTION ON A CHORDATE HISTOCOMPATIBILITY LOCUS

Marie L. Nydam,[1,2,3] Alyssa A. Taylor,[3] and Anthony W. De Tomaso[3]

[1]*Division of Science and Mathematics, Centre College, Danville, Kentucky 40422*

[2]*E-mail: marie.nydam@centre.edu*

[3]*Department of Molecular, Cellular, and Developmental Biology, University of California Santa Barbara, Santa Barbara, California 93106*

Received June 7, 2011
Accepted July 31, 2012

Allorecognition is the ability of an organism to differentiate self or close relatives from unrelated individuals. The best known applications of allorecognition are the prevention of inbreeding in hermaphroditic species (e.g., the self-incompatibility [SI] systems in plants), the vertebrate immune response to foreign antigens mediated by MHC loci, and somatic fusion, where two genetically independent individuals physically join to become a chimera. In the few model systems where the loci governing allorecognition outcomes have been identified, the corresponding proteins have exhibited exceptional polymorphism. But information about the evolution of this polymorphism outside MHC is limited. We address this subject in the ascidian *Botryllus schlosseri*, where allorecognition outcomes are determined by a single locus, called FuHC (Fusion/HistoCompatibility). Molecular variation in FuHC is distributed almost entirely within populations, with very little evidence for differentiation among different populations. Mutation plays a larger role than recombination in the creation of FuHC polymorphism. A selection statistic, neutrality tests, and distribution of variation within and among different populations all provide evidence for selection acting on FuHC, but are not in agreement as to whether the selection is balancing or directional.

KEY WORDS: **Allorecognition, ascidian, *Botryllus schlosseri*, FuHC.**

Allorecognition is the ability of an organism to differentiate self or close relatives from unrelated individuals. The best-known applications of allorecognition are the self-incompatibility (SI) systems in plants, vertebrate immune response to foreign antigens mediated by MHC loci, and fusion, where two genetically independent individuals physically join to become a single individual. Fusion occurs in a wide diversity of organisms (Buss 1982), including anemones (Mercier et al. 2011), ascidians (Oka and Watanabe 1957, Schmidt 1982, Bishop and Sommerfeldt 1999), bryozoans (Hughes et al. 2004), cellular slime molds (Shaulsky et al. 2007), corals (Hidaka et al. 1997), fungi (Glass et al. 2000), hydroids (Grosberg et al. 1996), plasmodial slime molds (Clark 2003), red algae (Santelices et al. 1999), and sponges (Wilson 1907). Effective allorecognition systems are critical to the survival of organisms: the SI locus prevents inbreeding depression, T-lymphocytes educated by MHC molecules protect vertebrates against pathogens, and fusing to a closely related individual can provide competitive and reproductive advantages where space is limited and reproductive output is based on the size of the organism (Buss 1982).

In the few systems where the loci governing allorecognition outcomes have been identified, the corresponding proteins have often exhibited exceptional polymorphism. In the clover *Trifolium pratense,* up to 193 S-alleles in the SI system were identified (Lawrence 1996). In human MHC, Class I A, Class I B, and Class II DRB1 loci have 1365, 1898, and 891 corresponding proteins, respectively (IMGT/HLA database: http://www.ebi.ac.uk/imgt/hla/stats.html, accessed June 26, 2012). In the colonial hydroid *Hydractinia symbiolongicarpus,* 35 alleles of the *alr2* allorecognition locus were sequenced

EVOLUTION INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

from 18 colonies (Rosengarten et al. 2010). But information about polymorphism in allorecognition loci, aside from MHC, is limited, including how it is created and maintained.

Regarding the processes that create MHC variation, some studies support a role for point mutations (Takahata et al. 1992), others for intragenic recombination and/or gene conversion (Richman et al. 2003, von Salome and Kukkonen 2008), although recombinational processes enhance existing variation (Parham and Ohta 1996).

Both neutral (i.e., genetic drift) (e.g., Miller et al. 2010) and selective forces (e.g., Aguilar et al. 2004) have been evoked for the maintenance of MHC polymorphism. In systems where selection has been detected, no consensus exists on the relative roles of three main types of balancing selection (spatially/temporally variable, heterosis, and negative frequency-dependent) (Hedrick 1999). In contrast, polymorphism at SI loci is generally thought to be maintained by negative frequency-dependent balancing selection (Wright 1939; Sato et al. 2002; Edh et al. 2009; Guo et al. 2011).

The well-characterized allorecognition system in the colonial ascidian *Botryllus schlosseri* provides an excellent opportunity to address questions of creation and maintenance of polymorphism in allorecognition loci. In *B. schlosseri*, the polymorphisms of a single gene called FuHC (Fusion/HistoCompatibility) determine 100% of histocompatibility outcomes between interacting colonies: fusion occurs if the colonies share one or more FuHC alleles (De Tomaso et al. 2005).

FuHC comprises 31 exons and the coding region is approximately 3 kb. Although RACE (Rapid Amplification of cDNA Ends) confirms that all 31 exons are contained in a single mRNA transcript, this transcript is not found in either mRNAseq or Northern Blot data (De Tomaso et al. unpubl. data). Instead, FuHC is expressed as two separate transcripts. The first transcript encodes a secreted form of the protein Exons 1–17 (Exons 15–17 are added by alternative splicing) and contains a predicted signal sequence (Exon 1) and two sets of epidermal growth factor (EGF) repeats (Exons 8/9, 15/16) (De Tomaso et al. 2005). The second transcript contains Exons 18–29/31 (Exon 30 is almost always spliced out) and codes two predicted immunoglobulin (Ig) domains (Exons 18–20 and Exons 21–22) and a transmembrane (TM) domain (Exon 27). This information leads us to hypothesize that FuHC is actually an operon: two genes sharing a promoter. Operons are abundant in the genome of another ascidian, Type A *Ciona intestinalis* (Zeller 2010). When we use the term "FuHC locus," we are describing a physical location on a chromosome.

FuHC is expressed as two transcripts encoding two putative protein products that likely perform disparate functions and experience different selection pressures; it may even be two genes. We therefore amplified, sequenced, and analyzed each transcript separately.

As in other allorecognition loci, FuHC exhibits high polymorphism: preliminary studies of FuHC variation found 18 FuHC alleles from 10 individuals in a single population (De Tomaso et al. 2005). *Botryllus schlosseri* populations are predicted to contain up to 300 FuHC alleles (Grosberg and Quinn 1986, Rinkevich et al. 1995). In contrast to the MHC, the single locus FuHC controls allorecognition in *B. schlosseri* by itself, and no duplications or pseudogenes exist. Studies involving FuHC therefore avoid the complications and limitations imposed by the complexity of the polygenic MHC.

In the current study, we address two aspects of FuHC evolution: (1) which evolutionary forces create variation, and (2) which evolutionary forces maintain variation. Investigating these questions will increase our limited understanding of the evolution of allorecognition loci.

Materials and Methods

SAMPLING

A total of 20–40 colonies were collected from floating docks in each of six populations in 2009 and 2010: Falmouth, MA; Quissett, MA; Sandwich, MA; Monterey, CA; Santa Barbara, CA; and Seattle, WA. Single systems were dissected from individual colonies and flash frozen with liquid nitrogen and stored at −80°C. *Botryllus schlosseri* has recently been shown to be a complex of five genetically distinct and divergent clades (Bock et al. 2012). All our samples are Clade A. Clade A is native to the Mediterranean and has spread throughout the Atlantic and Pacific Oceans. Populations from the West Coast of the United States (Monterey, CA; Santa Barbara, CA; and Seattle, WA) are likely derived from the Western Pacific in two separate introduction events (Lejeusne et al. 2010). Populations from the East Coast of the United States (Falmouth, MA; Quisset, MA; and Sandwich, MA) come from Europe, in at least two introduction events (Lejeusne et al. 2010).

AMPLIFICATION AND SEQUENCING

FuHC

Total RNA was extracted from frozen tissue using the NucleoSpin Nucleic Acid and Protein Purification Kit (Macherey-Nagel). This RNA was used to synthesize single-stranded cDNA using SuperScript III reverse transcriptase (Invitrogen) and an oligo (dT) primer. Fivefold dilutions of the single-stranded cDNA were then PCR amplified with TRsa and TS-PCR primers. The resulting PCR product was diluted 50-fold and used as the template for PCR amplification. Exons 1–14 of the secreted transcript (Exons 1–17) and Exons 18–29/31 were amplified and sequenced separately. PCR primers for (Exons 1–14) are WholeIgF2 (5′ GAAATGTTGCTGAAAATATTCTGTCTT

3′) and 14–18soeR (5′ TACTTCAAGTCGACAGTTCCAATCAACGTA 3′). PCR primers for Exons 18–29/31 are WholeIgR2 (5′ GATACTTGGCTCTCGCCTTGATCTT 3′) and 14–18soeF (5′ TACGTTGATTGGAACTGTCGACTTGAAGTA 3′). Cycling conditions for Exons 1–14 were 35x (95°C for 30 s, 51°C for 30 s, 72°C for 2 min), 72°C for 5 min, and for Exons 18–29/31 were 35x (95°C for 30 s, 55°C for 30 s, 72°C for 2 min), 72°C for 5 min. PCR amplification was performed in a 20-μl total reaction volume with 13.6 μl of H_2O, 4 μl of 5x HF buffer (Finnzymes), 0.2 mM dNTPs, 0.6 μl of 100% DMSO, 0.3333 μM of each primer, 0.02U/μl of Phusion Polymerase (Finnzymes), and 2 μl of template DNA. PCR products were cloned using the pGEM ® -T kit and up to 10 clones were sequenced. Colony PCR products were incubated with 0.25 μl each of Exonuclease I and Shrimp Antarctic Phosphatase at 37°C for 30 min, followed by 90°C for 10 min prior to sequencing.

Purified PCR products were sequenced with a Big Dye Terminator Cycle sequencing kit and a 96 capillary 3730xl DNA Analyzer (Applied Biosystems) at the UC Berkeley Sequencing Facility. All sequences have been submitted to GenBank (FuHC Exons 1–14: JN082923-JN083035, FuHC Exons 18–29/31: JN083036-JN083147). Sequences were edited, trimmed, and aligned with Aligner (CodonCode Corporation, Dedham, MA).

Housekeeping genes

We amplified 12 nuclear housekeeping genes to determine whether the pattern of population structure and the values of neutrality tests seen at FuHC were specific to that locus. Significant negative values of neutrality tests could be due to selective or demographic processes (e.g., recent population growth). But demographic processes would affect all genes, not just those involved in allorecognition. Two of the 12 nuclear loci were found in GenBank (adult-type muscle actin 2, Accession no. FN178504.1 and vasa, Accession no. FJ890989.1) and the other 10 were located in our *B. schlosseri* EST database (40S ribosomal protein 3A, 60S ribosomal protein L6, 60S ribosomal protein L8, 60S ribosomal protein L10, 60S ribosomal protein L13, heat shock cognate 71 kDa protein, cytoplasmic actin 2, ADP/ATP translocase 3, heat shock protein HSP-90 beta, and vigilin).

Template for PCR amplification was generated as described above for FuHC. Primers and thermocycling conditions for each gene are available from the authors. vasa PCR products were cloned as described for FuHC. The PCR products of the other nuclear loci were sequenced directly. PCR products were incubated with 0.25 μl each of Exonuclease I and Shrimp Antarctic Phosphatase at 37°C for 30 min, followed by 90°C for 10 min.

Purified PCR products were sequenced with a Big Dye Terminator Cycle sequencing kit and a 96 capillary 3730xl DNA Analyzer (Applied Biosystems) at the UC Berkeley Sequencing Facility. Sequences that were obtained by direct sequencing of PCR products (all nuclear sequences minus vasa) were phased in DnaSP 5.10.01 (Librado and Rozas 2009). All sequences have been submitted to GenBank (40S ribosomal protein 3A: JQ596880-JQ596936, 60S ribosomal protein L6: JQ596937-JQ597084, 60S ribosomal protein L8: JQ597085-JQ597174, 60S ribosomal protein L10: JQ597175-JQ597294, 60S ribosomal protein L13: JQ597595-JQ597716, heat shock cognate 71 kDa protein: JQ597295-JQ597430, cytoplasmic actin 2: JQ597431-JQ597548, ADP/ATP translocase 3: JQ597549-JQ597594, heat shock protein HSP-90 beta: JQ597717-JQ597826, adult-type muscle actin 2: JQ597827-JQ597974, vasa: JN083304-JN083376, vigilin: JQ597975-JQ598070). Sequences were edited, trimmed, and aligned with Aligner (CodonCode Corporation, Dedham, MA).

Amino acid diversity across FuHC

We employed the program DIVAA (Rodi et al. 2004) to calculate the amino acid variation across FuHC. Amino acid alignments for Exons 1–14 and Exons 18–29/31 were generated and analyzed separately. A diversity value of 1 for a particular position means that any amino acid is as likely as any other to be present at that position. A diversity value of 0.05 is consistent with 100% conservation at that site (1/20 possible amino acids). We also determined the number of protein alleles in our data set for Exons 1–14 and Exons 18–29/31.

Recombination

Intragenic recombination in each population for FuHC Exons 1–14 and Exons 18–29/31 was assessed by calculating R_m, the minimum number of recombination events in DnaSP 5.10.01 (Librado and Rozas 2009) and the correlation between physical distance and three measures of linkage disequilibrium: r^2, D' and G4 in program permute (Wilson and McVean 2006).

Levels of recombination (ρ) and associated 95% highest posterior density (HPD) regions across FuHC were estimated using the program omegaMap 0.5 (Wilson and McVean 2006). omegaMap runs were carried out using the resources of the Computational Biology Service Unit at Cornell University, which is partially funded by the Microsoft Corporation. We chose 250,000 iterations for each run, with thinning set to 1000. We used an improper inverse distribution for μ (rate of synonymous transversion), and κ (transition–transversion ratio), and an inverse distribution for ω (selection parameter) and ρ (recombination rate). Initial parameter values for μ and κ were 0.1, and 3.0, respectively. ω and ρ priors were set between 0.01 and 100. An independent model was used for ρ, so that recombination values were allowed to vary across sites. The number of iterations discarded as burn-in varied across runs, but was determined by plotting the traces of μ and κ; iterations affected by the starting value of the parameter

were discarded. Two independent runs were conducted for each population. These two runs were combined in all cases, after it was determined that the mean and 95% HPD regions for each parameter in the two runs matched closely.

Creation of FuHC polymorphism: Relative roles of mutation and recombination

The relative contributions of mutation and recombination in generating polymorphism in FuHC were determined by estimating θ_w per site and R between adjacent sites in DnaSP 5.10.01 (Librado and Rozas 2009). R between adjacent sites is calculated by obtaining R per gene. R per gene $= 4N * r$ (Hudson 1987), where N is the population size and r is the recombination rate per sequence. R per gene is then divided by D (the average nucleotide distance for the gene) to obtain R between adjacent sites. A ratio of $\theta/R = 1$ signifies an equal contribution of mutation and recombination, >1 a larger role for mutation, and <1 a larger role for recombination.

Maintenance of polymorphism: Selection statistics

Selection statistic values and associated 95% HPD regions across each FuHC transcript were estimated using the program omegaMap 0.5 (Wilson and McVean 2006). Settings were identical to the recombination analyses (see above), except that an independent model was used for ω. We also calculated the posterior probability of selection per codon across the each transcript. Codons with a >95% posterior probability of selection in at least three of the six populations were selected for further analyses. Exons that contained clusters (>2) of these codons were identified; Mann–Whitney U-tests in R 2.15.0 were performed on these exons to determine if they had higher selection statistic values than the rest of the corresponding transcript (Exons 1–14: Exon 5, Exon 6 or Exons 18–29/31: Exon 20, Exon 27). Selection statistic values for exons containing predicted domains (Exons 8–9, 18–20, 21–22, 27; De Tomaso et al. 2005) were compared with selection statistic values in the rest of the corresponding transcript (Exons 1–14 or Exons 18–29/31) using a one-tailed Mann–Whitney U-tests as above.

Maintenance of polymorphism: Neutrality tests

The summary statistics θ, π, number of haplotypes, and haplotype diversity were calculated in DnaSP 5.10.01 (Librado and Rozas 2009) for FuHC Exons 1–14, 18–29/31, and all housekeeping genes. For FuHC Exons 1–14 and 18–29/31, we tested whether π based on nonsynonymous sites is higher than π based on synonymous sites, using a one-tailed t-test in R 2.15.0.

We also employed Tajima's D (Tajima 1989) test statistic for FuHC and all housekeeping genes. Tajima's D statistics were calculated for all sites together, and for nonsynonymous and synonymous sites separately. We only calculated D values based on nonsynonymous sites for housekeeping genes if the alignment contained greater than six polymorphic nonsynonymous sites.

We calculated Tajima's D for four different sampling schemes to test for a "pooling effect" (Stadler et al. 2009). This effect occurs when population-level sampling (aka "local" sampling) yields site-frequency spectra that are skewed toward intermediate frequency alleles when compared to spectra obtained from grouping alleles from multiple populations into the same sample (aka "pooled" sampling) (Stadler et al. 2009; Cutter et al. 2012). The four sampling schemes were as follows: (1) all alleles pooled into a single sample (pooled all), (2) all East Coast alleles pooled into one sample, and all West Coast alleles pooled into a second sample (pooled East/West Coast), (3) each of the six populations as a separate sample (local), and (4) "scattered" sampling. For scattered sampling, we assembled 1000 samples, each composed of one allele drawn randomly from each population (Cutter et al. 2012). We constructed these 1000 samples using the ape package in R 2.15.0 (Paradis 2012).

Tajima's D for pooled all, pooled East/West Coast, and local sampling schemes was calculated in DnaSP 5.10.01 (Librado and Rozas 2009). Statistical significance of D for these sampling schemes was determined using 1000 coalescent simulations in DnaSP 5.10.01 (Librado and Rozas 2009). We performed two sets of coalescent simulations based on θ and segregating sites. Estimates of per gene recombination (R) for each population were made in DnaSP 5.10.01 (Librado and Rozas 2009) and were then imported into the simulations. Tajima's D values for scattered samples were calculated using the pegas package in R 2.15.0 (Paradis 2010).

Information about Nm (migration rate) provides insight into the pooling effect (Stadler et al. 2009). Values of Nm were estimated from F_{ST} values for all housekeeping genes and both FuHC transcripts using DnaSP 5.10.01 (Librado and Rozas 2009). Mean Nm values were compared between housekeeping genes with negative and positive D values using a two-sample t-test in R 2.15.0.

We performed sliding window analyses for each population for FuHC Exons 1–14 and Exons 18–29/31 in DnaSP 5.10.01 (Librado and Rozas 2009). The window size was 100 bp and the step size was 25.

Maintenance of polymorphism: Distribution of polymorphism within and among different populations

We characterized population structure within *B. schlosseri* for FuHC Exons 1–14, Exons 18–29/31, and all housekeeping genes using an analysis of molecular variance (AMOVA), fixation indices (F_{CT}, F_{SC}, and F_{ST}), and pairwise F_{ST} values between all populations in Arlequin 3.5.1.2 (Excoffier and Lischer 2010). For FuHC transcripts, we also performed AMOVA/F-statistic calculations for synonymous and nonsynonymous sites separately. The two groups in these analyses were U.S. East Coast (Falmouth,

MA; Quissett, MA; and Sandwich, MA) and U.S. West Coast (Monterey, CA; Santa Barbara, CA; Seattle, WA).

Results

AMINO ACID DIVERSITY ACROSS FUHC

Amino acid diversity across FuHC is presented in Figure S1. The two transcripts (Exons 1–14 and Exons 18–29/31) have striking differences in both mean diversity and standard deviation around the mean. The Exon 18–29/31 transcript has lower mean diversity and lower standard deviation around the mean than Exon 1–14. Amino acids in Exons 18–29/31 can be completely conserved (diversity value = 0.05) or nearly so; this is never the case in Exons 1–14. We sequenced 77 individuals for Exons 1–14 and recovered 42 protein alleles; for Exons 18–29/31, 76 individuals and 44 protein alleles.

RECOMBINATION

Intragenic recombination has occurred within FuHC. Minimum number of recombination events (R_m) was 9 for Exons 1–14 and 7 for Exons 18–29/31. Significant negative correlations between physical distance and all three measures of linkage disequilibrium exist for Exons 1–14 and Exons 18–29/31 ($P = 0.001$ in all cases). No recombination peaks or troughs were identified across FuHC in any of the six populations: recombination values were similar across each transcript. Figure S2 shows recombination across FuHC for a representative population: Monterey, CA.

CREATION OF FUHC POLYMORPHISM: RELATIVE ROLES OF MUTATION AND RECOMBINATION

Mutation plays a larger role in the creation of FuHC polymorphism than does recombination; for both Exons 1–14 and Exons 18–29/31, $\theta/R > 1$. For Exons 1–14, θ per site = 0.05, R between adjacent sites = 0.01, and the ratio of $\theta/R = 3.35$. For Exons 18–29/31, θ per site = 0.05, R between adjacent sites = 0.02, and the ratio of $\theta/R = 2.78$.

MAINTENANCE OF POLYMORPHISM: SELECTION STATISTICS

Codons with selection statistic values >1 are spread throughout Exons 1–14 and Exons 18–29/31, but many of these codons have 95% HPD regions that include 1. Figure S3 shows the distribution of selection statistic values across Exons 1–14 and Exons 18–29/31 for a typical population: Monterey. The number of codons that have >95% posterior probability of selection ranges from 34 (Monterey, CA) to 45 (Santa Barbara, CA and Seattle WA). Twenty-four of these codons are present in three or more of the six populations. These 24 codons are concentrated in Exons 1–14 (Fig. S4). There are three clusters of these codons: Codons 117, 119, 130, 147, 150, 171, 179, 182, and 185 (Exons 5 and 6), Codons 575 and 576 (Exon 20), and Codons 851, 854, and 858 (Exon 27). Exon 5 had significantly higher selection statistic values than the rest of the Exon 1–14 transcript in Falmouth, MA (U value = 14,408, $P = 0.004$); Quissett, MA (U value = 14,804, $P = 0.001$); Sandwich, MA (16,392, $P = 7.9 \times 10^{-7}$); and Seattle, WA (U value = 13,574, $P = 0.04$) (Table 1). Exon 6 had significantly higher selection statistic values than the rest of the Exon 1–14 transcript in Falmouth, MA (U value = 9917, $P = 0.015$), and Sandwich, MA (10,123, $P = 0.007$). Exon 20 had significantly higher selection statistic values than the rest of the Exon 18–29/31 in the Quissett, MA population only (U value = 14.058, $P = 0.0001$). Exon 27 (also a predicted TM domain) did not have significantly higher selection statistic values than the rest of the Exon 18–29/31 transcript for any populations.

We also tested whether exons containing predicted domains had significantly higher selection statistic values than the rest of the corresponding transcript (Exons 1–14 or Exons 18–29/31). Exons 8 and 9 (predicted EGF domain), Exons 18–20 (predicted Ig domain), Exons 21–22 (predicted Ig domain), and Exon 27 (predicted TM domain) did not show any significant U-test P-values for any population, excepting one population for Exons 21–22 (Seattle, WA).

MAINTENANCE OF POLYMORPHISM: NEUTRALITY TESTS

FuHC

θ, π, number of haplotypes, and haplotype diversity for each population and transcript (Exons 1–14 and Exons 18–29/31) are shown in Table S1. θ and π values for Exons 1–14 averaged across all populations are as follows: θ (per site, all sites): 0.037, π (all sites): 0.032, π (synonymous sites): 0.02, π (nonsynonymous sites): 0.036. θ and π values for Exons 18–29/31 averaged across all populations are as follows: θ (per site, all sites): 0.029, π (all sites): 0.025, π (synonymous sites): 0.063, π (nonsynonymous sites): 0.020. π (nonsynonymous sites) is greater than π (synonymous sites) for Exons 1–14 ($P < 0.001$) but not for Exons 18–29/31 ($P > 0.05$).

Mean Tajima's D values for each FuHC transcript (Exons 1–14 and Exons 18–29/31) for each of the sampling schemes are shown in Table 2. When synonymous and nonsynonymous sites are considered together, both FuHC transcripts exhibit a distinct pooling effect. Scattered sampling gives the lowest D value, then pooled samples, then local samples. Regardless of the sampling scheme, all D values are negative for both FuHC transcripts.

Table 3 includes individual values for East Coast/West Coast and local samples, as well as statistical significance of D values based on P values obtained from coalescent simulations for both FuHC transcripts. Statistical significance was not determined for

Table 1. Mann–Whitney *U* statistics for testing whether exons of interest have higher omega values than the rest of the transcript. Asterisks denote statistical significance at $\alpha = 0.05$.

Population	Exon 5	Exon 6	Exons 8 and 9	Exons 18–20	Exon 20	Exons 21–22	Exon 27
Falmouth, MA	14,408*	9917*	11,290	18,678	10,379	9825	10,899
Quissett, MA	14,804*	8724	12,833	11,921	14,058*	10,392	10,818
Sandwich, MA	16,392*	10,123*	9400	18,542	10,540	10,111	10,145
Monterey, CA	10,449	8025	10,676	17,547	11,889	11,406	10,399
Santa Barbara, CA	12,997	9093	10,221	18,416	11,312	11,035	11,453
Seattle, WA	13,574*	8210	12,641	14,672	11,577	13,366*	9042

Table 2. Tajima's *D* neutrality test for FuHC Exons 1–14, FuHC Exons 18–29/31, and 12 housekeeping genes.

	D_{Taj}: Pooled all	D_{Taj}: Pooled East/West[a]	D_{Taj}: Local[b]	D_{Taj}: Scattered[c]
FuHC Exons 1–14, all sites	−1.56	−1.05	−0.67	−2.69
FuHC Exons 1–14, NS only	−1.16	−0.84	−0.58	−0.45
FuHC Exons 1–14, S only	−1.68	−1.23	−0.92	−0.47
FuHC Exons 18–29/31, all sites	−1.41	−0.97	−0.60	−2.09
FuHC Exons 18–29/31, NS only	−1.41	−1.33	−0.85	−0.04
FuHC Exons 18–29/31, S only	−1.02	−0.46	−0.27	−0.02
Housekeeping genes, all sites	0.06	0.20	0.02	−0.13
Housekeeping genes, S only	0.13	0.29	0.05	0.02

All sites: synonymous and nonsynonymous sites. NS = nonsynonymous sites. S = synonymous sites. NS only values for housekeeping genes are not reported as only 3/12 housekeeping gene alignments contained polymorphic nonsynonymous sites.

[a]Mean of D_{Taj} values for East Coast and West Coast.

[b]Mean of D_{Taj} values for all six populations.

[c]Mean of D_{Taj} values for all 1000 samples.

scattered samples, so these values are excluded. Simulations given θ and segregating sites give similar *P* values (and always with the same qualitative result), so only results from the simulations given θ will be shown.

Exons 1–14 display a similar pattern in significance for both synonymous and nonsynonymous sites regardless of the sampling scheme. For instance, pooled all, pooled East Coast, Sandwich, MA, and Monterey, CA *D* values are significantly negative for both synonymous and nonsynonymous sites. Where values based on synonymous and nonsynonymous sites do not have the same statistical significance, the values based on synonymous sites are always significantly negative and the nonsynonymous values are not. Additionally, no clear pattern emerges regarding the relationship between *D* values based on all sites and *D* values based on synonymous sites.

On the other hand, Exons 18–29/31 are significantly negative only when considering nonsynonymous sites (e.g., pooled East Coast; pooled West Coast; Quissett, MA; Sandwich, MA; and Santa Barbara, CA). *D* values based on synonymous sites are not statistically different from zero for any of the populations. Also, *D* values based on all sites are consistently less than values based on synonymous sites.

Sliding window Tajima's *D* analyses of Exons 1–14 and Exons 18–29/31 did not generally locate windows that had Tajima's *D* values that were significantly different from zero, although both positive and negative windows were often seen in a single population (data not shown).

Housekeeping genes

The mean *D* values across all housekeeping genes for all sites and synonymous sites only for all sampling schemes can be found in Table 2. Mean *D* values based on nonsynonymous sites are not shown, as only three of the 12 housekeeping genes have more than six nonsynonymous polymorphic substitutions. In contrast to the FuHC transcripts, we did not detect a pooling effect, as values were very similar between sampling regimes. Regardless of the sampling scheme used for housekeeping genes, they seem to be evolving very close to neutrality (i.e., no evidence for demographic events).

Table S2 includes individual values for East Coast/West Coast and local samples, as well as statistical significance of *D* values based on *P* values obtained from coalescent simulations for both FuHC transcripts. Statistical significance was not determined for scattered samples, so these values are excluded. Eleven

Table 3. Tajima's *D* values, associated 95% confidence intervals and significance at $\alpha = 0.05$ for FuHC Exons 1–14 and Exons 18–29/31, based on pooled all, pooled East/West Coast, and local sampling schemes.

FuHC Exons 1–14			
Pooled all	*D* (95% CI): All sites	*D* (95% CI): Nonsynon. sites	*D* (95% CI): Synon. sites
	−1.56 (−1.16 to 1.05)*	−1.16 (−1.23 to 1.04)*	−1.68 (−1.29 to 1.16)*
Pooled East/West Coast	*D* (95% CI): All sites	*D* (95% CI): Nonsynon. sites	*D* (95% CI): Synon. sites
East Coast	−1.49 (−1.26 to 1.14)*	−1.29 (−1.29 to 1.17)*	−1.16 (−1.22 to 1.08)*
West Coast	−0.60 (−1.11 to 1.01)	−0.39 (−1.07 to 1.04)	−1.299 (−1.13 to 1.08)*
Local Population	*D* (95% CI): All sites	*D* (95% CI): Nonsynon. sites	*D* (95% CI): Synon. sites
Falmouth, MA	−1.24 (−1.52 to 1.38)	−1.12 (−1.49 to 1.39)	−1.76 (−1.49 to 1.39)*
Quissett, MA	−0.64 (−0.69 to 0.69)*	−0.48 (−1.19 to 1.07)	−0.81 (−1.21 to 1.08)
Sandwich, MA	−0.58 (−1.19 to 1.11)	−1.17 (−1.14 to 1.02)*	−1.33 (−1.05 to 1.06)*
Monterey, CA	−0.02 (−10.35 to 0.92)	−0.61 (0.62 to 0.73)*	−0.63 (0.67 to 0.70)*
Santa Barbara, CA	−0.22 (−1.53 to 1.42)	0.09 (−1.10 to 1.00)	−0.61 (−1.01 to 0.91)
Seattle, WA	−1.32 (−1.02 to 1.00)*	−0.18 (−1.48 to 1.41)	−0.40 (−1.43 to 1.50)
FuHC Exons 18–29/31			
Pooled all	*D* (95% CI): All sites	*D* (95% CI): Nonsynon. sites	*D* (95% CI): Synon. sites
	−1.41 (−1.11 to 1.18)*	−1.41 (−1.23 to 1.18)*	−1.02 (−1.17 to 1.14)*
Pooled East/West Coast	*D* (95% CI): All sites	*D* (95% CI): Nonsynon. sites	*D* (95% CI): Synon. sites
East Coast	−1.21 (−1.23 to 1.25)*	−1.51 (−1.26 to 1.18)*	−0.70 (−1.23 to 1.28)
West Coast	−0.72 (−1.22 to 1.05)	−1.14 (−1.22 to 1.10)*	−0.22 (−1.18 to 1.19)
Local Population	*D* (95% CI): All sites	*D* (95% CI): Nonsynon. sites	*D* (95% CI): Synon. sites
Falmouth, MA	−0.59 (−0.93 to 0.84)	−0.54 (−0.93 to 0.80)	−0.31 (−0.83 to 0.83)
Quissett, MA	0.1 (−0.78 to 0.70)	−1.27 (−1.44 to 1.38)*	−0.81 (−1.44 to 1.38)
Sandwich, MA	−1.1 (−1.58 to 1.33)	−1.35 (−1.09 to 1.05)*	−0.31 (−1.13 to 1.03)
Monterey, CA	−0.51 (−0.97 to 0.96)	−0.15 (−0.78 to 0.78)	0.35 (−0.73 to 0.71)
Santa Barbara, CA	−0.66 (−1.51 to 1.53)	−1.20 (−0.96 to 0.93)*	−0.09 (−0.88 to 0.93)
Seattle, WA	−0.81 (−1.16 to 1.09)	−0.59 (−1.62 to 1.52)	−0.46 (−1.61 to 1.57)

*statistical significance at $\alpha = 0.05$

of the 12 genes have no populations with *D* values statistically different from zero. One of the genes (60S ribosomal protein L10) has a single population with a *D* value statistically less than zero. In stark contrast to FuHC, we see very few significant values for the housekeeping loci. For 60S ribosomal protein L10, one population is significant. None of the other housekeeping loci have populations with significant values.

We also noted that nearly all *D* values across populations were negative for the FuHC transcripts. For the housekeeping genes, only 60S ribosomal protein L6 and adult-type muscle actin 2 show a pattern of consistent negative values across populations for Tajima's neutrality test. Seven genes (40S ribosomal protein 3A, 60S ribosomal protein L8, 60S ribosomal protein L10, ADP/ATP translocase, HSP 90 beta, heat shock cognate 71 kDa protein, and vasa) show no trend toward positive or negative values across populations. 60S ribosomal protein L6 and adult-type muscle actin 2 are negative across populations for Tajima's *D*. 60S ribosomal protein L13 shows a pattern of positive values across populations.

To compare housekeeping genes and FuHC, we calculated mean Tajima's *D* values (across all six populations, local sampling) for housekeeping genes. For each FuHC transcript, we determined mean Tajima's *D* values (across East Coast and West Coast, pooled East/West Coast sampling) for synonymous and nonsynonymous sites separately. We used the pooled samples because we determined that a pooling effect is skewing local site-frequency spectra toward intermediate frequencies. Figure 1 shows that mean *D* values for Exons 1–14 nonsynonymous sites, Exons 1–14 synonymous sites, and Exons 18–39/31 nonsynonymous sites are substantially lower than values for all but 2 of the 12 housekeeping genes (60S ribosomal protein L6 and adult-type muscle actin 2).

MAINTENANCE OF POLYMORPHISM: DISTRIBUTION OF POLYMORPHISM WITHIN AND AMONG DIFFERENT POPULATIONS

FuHC

Analyses based on AMOVA suggest that most variation is within populations, a small amount of variation is found among population within groups, and no variation is found among groups (U.S. East and West Coasts) (Table 4). For Exons 1–14, F_{CT}, F_{SC},

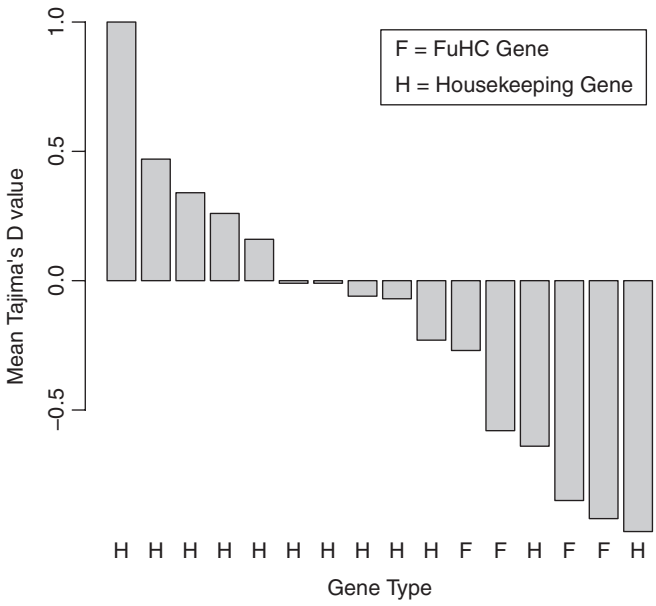

Figure 1. Comparison of mean Tajima's *D* between FuHC and housekeeping genes. FuHC genes are labeled "F," and housekeeping genes are labeled "H." There are four FuHC values: Exons 1–14 nonsynonymous sites only (NonSyn), Exons 1–14 synonymous sites only (Syn), Exons 18–29/31 nonsynonymous sites only (NonSyn), and Exons 18–29/31 synonymous sites only (Syn). Mean Tajima's *D* values for housekeeping genes were obtained by averaging across all six populations (Falmouth, MA; Quissett, MA; Sandwich, MA; Monterey, CA; Santa Barbara, CA; and Seattle, WA). Mean Tajima's *D* values for FuHC transcripts were obtained by averaging across East Coast and West Coast pooled samples.

and F_{ST} are not significant ($P > 0.05$), and only the Falmouth, MA versus Sandwich, MA pairwise F_{ST} is significant. For Exons 18–29/31, F_{SC} and F_{ST} are significant. Only the five pairwise comparisons involving Santa Barbara, CA are significant; the Santa Barbara, CA population is likely responsible for the 9.3% variation among population within groups and significant F_{SC} and F_{ST} values.

When synonymous and nonsynonymous sites are analyzed separately, the results are similar to those obtained from all sites analyses. Parsing out synonymous and nonsynonymous sites reveals more population differentiation for Exons 1–14 at nonsynonymous sites, as evidenced by a slight (2.4%) variation among populations within groups compared to the percentage seen when nonsynonymous and synonymous sites are taken together, and a newly significant F_{ST} value. No differences in any of the *F*-statistics were seen in Exons 18–29/31 between nonsynonymous and synonymous sites analyzed separately and together. Both FuHC transcripts have higher than 90% of the variation within population variation and less than 10% among different populations/among groups regardless of whether the analyses were performed on synonymous/nonsynonymous sites separately or together.

Housekeeping genes

The housekeeping genes show a pattern that contrasts with FuHC. Figure 2 shows that all housekeeping loci have a lower percentage of variation within populations than either FuHC transcript, and in many cases a substantially lower percentage. All housekeeping genes have a substantially higher percentage of variation among groups than FuHC although F_{CT} is not significant for any locus (Table S3). All housekeeping genes have a substantially higher percentage of variation among populations within groups than FuHC, and F_{SC} is significant for all loci but 1 (Table S3). Significant differentiation exists among populations among groups for housekeeping genes: overall F_{ST} is significant for all loci, and a majority of pairwise F_{ST} values are significant.

Discussion

AMINO ACID DIVERSITY ACROSS FUHC

Amino acid diversity differs substantially between Exons 1–14 and Exons 18–29/31. Given that these gene regions correspond to two separate transcripts, we hypothesize that they encode separate proteins. These proteins likely have separate functions, given the differences in amino acid conservation between them. Exons 1–14 encode most of the secreted form of the protein (De Tomaso et al. 2005), whereas the putative protein encoded by Exons 18–29/31 likely remains in the membrane. We do not yet understand the specific roles of the two putative proteins in the allorecognition reaction. We are currently performing in situ hybridizations to determine whether the two transcripts colocalize (De Tomaso et al., unpubl. data).

CREATION OF FUHC POLYMORPHISM: RELATIVE ROLES OF MUTATION AND RECOMBINATION

FuHC experiences a substantial amount of intragenic recombination, based on three independent measures: R_m, the correlation between physical distance and three measures of linkage disequilibrium, and levels of recombination across each transcript. However, this recombination clearly plays a smaller role in the creation of FuHC polymorphism than mutation, given that θ/R was much greater than one for both Exons 1–14 and Exons 18–29/31.

Recombination likely contributes to *alr2* polymorphism in *Hydractinia,* based on the discovery of chimeric alleles having regions characteristic of two distinct types of structural polymorphism (Rosengarten et al. 2010), but the relative contributions of mutation and recombination to allelic diversity were not assessed. The relative roles of mutation and recombination in generating diversity at MHC loci remain the subject of debate. Some workers

Table 4. AMOVA, fixation indices, and pairwise F_{ST} values for FuHC. F_{CT}, F_{SC}, and F_{ST} are the *F*-statistics. Asterisks denote significance at $\alpha = 0.05$.

Exons 1–14						
Source of variation	df	Sum of squares	Variance	% of variation	Fixation indices	
Among groups	1	163.14	−0.69	−0.53	F_{CT}: −0.01	
Among pops. within groups	4	767.78	3.61	2.77	F_{SC}: 0.03	
Within pops.	106	13,540.65	127.74	97.76	F_{ST}: 0.02	
Pairwise F_{ST}						
	Falmouth	Quissett	Sandwich	Monterey	Santa Barbara	Seattle
Falmouth	0	0.08	0.01*	−0.05	0.04	0.01
Quissett		0	−0.02	0.04	−0.02	0.02
Sandwich			0	0.07	−0.02	0.06
Monterey				0	−0.01	−0.02
Santa Barbara					0	0
Seattle						0
Exons 18–29/31						
Source of variation	df	Sum of squares	Variance	% of variation	Fixation indices	
Among groups	1	213.49	−2.16	−1.77	F_{CT}: −0.02	
Among pops. within groups	4	1270.47	11.34	9.31	F_{SC}: 0.09*	
Within pops.	106	11,936.59	112.61	92.46	F_{ST}: 0.08*	
Pairwise F_{ST}						
	Falmouth	Quissett	Sandwich	Monterey	Santa Barbara	Seattle
Falmouth	0	0.01	0.01	−0.04	0.34*	0.03
Quissett		0	−0.02	−0.03	0.21*	−0.01
Sandwich			0	−0.02	0.19*	−0.02
Monterey				0	0.27*	−0.004
Santa Barbara					0	0.26*
Seattle						0

have argued for the importance of point mutations (Takahata et al. 1992), others for intragenic recombination and/or gene conversion (Parham and Ohta 1996). Many MHC studies have found a greater role for recombination than mutation, concurring with the FuHC results presented here (e.g., Richman et al. 2003; Alcaide et al. 2008; von Salome and Kukkonen 2008). However, mutation and recombination likely work in concert to generate new alleles at allorecognition loci: recombination augments the existing diversity created by point mutations (Geliebter and Nathenson 1987).

MAINTENANCE OF FUHC POLYMORPHISM: SELECTION STATISTICS

The program omegaMap 0.5 pinpointed 24 codons throughout both transcripts that have a greater than 95% probability of the selection statistic >1. Four exon groups contained clusters of these selected sites: Exons 5, 6, 20, and 27. Exons 5, 6, and 20 had significantly higher selection statistic values than the rest of the corresponding transcript for a subset of populations (Exon 5: 4/6 populations, Exon 6: 3/6 populations, Exon 20: 1/6 populations; Table 1). No known domains were predicted for Exons 5 and 6 (De Tomaso et al. 2005); these exons, as a target of selection, will be the focal point of future mechanistic studies of FuHC. Exon 20 has a predicted Ig domain (along with Exons 18 and 19), but only 1/6 populations had significantly higher selection statistic values at this exon, and Exons 18–20 as a unit did not have significantly higher selection statistic values for any populations. Several other exons were predicted to have known domains contained within them (Exons 8 and 9: an EGF domain, Exons 18–20: an Ig domain, Exons 21–22: an Ig domain; De Tomaso et al. 2005). None of these exons have higher selection statistic values than the rest of the corresponding transcript (Table 1), and are not therefore likely targets for selection.

Although these analyses clearly show that selection is acting on FuHC, they cannot say whether balancing or directional selection is occurring. The selection statistic used here cannot discriminate between these two types of selection.

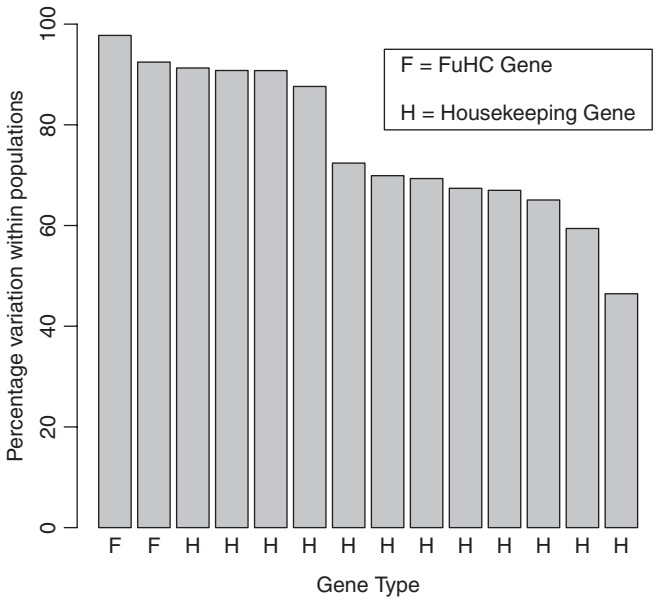

Figure 2. Comparison of percentage of variation found within populations between FuHC and housekeeping genes. FuHC genes are labeled "F," and housekeeping genes are labeled "H." These numbers were derived from AMOVA, which are presented in their entirety in Table 4 (for FuHC) and Table S3 (for housekeeping genes). Both FuHC transcripts have higher percentages of variation within populations than any of the housekeeping loci.

In the cellular slime mold *Dictyostelium discoideum,* the genes *lagB1* and *lagC1* are involved in kin recognition. Certain sections of these genes have d_N/d_S ratios > 1; the authors conclude that balancing selection is causing this pattern, given the extensive polymorphism at these loci (Benabentos et al. 2009). Nine codons in *alr2* of *Hydractinia symbiolongicarpus* have elevated d_N/d_S ratios; the majority are found in exon 2 (Rosengarten et al. 2010). The presence of 35 *alr2* alleles recovered from 36 individuals led the authors to conclude that negative frequency-dependent selection (a type of balancing selection where rare alleles are favored by selection) is occurring. At equilibrium, the alleles of a single locus subject to frequency-dependent selection are expected to be equally frequent (Grosberg 1988).

Neither of two mating-compatibility genes examined in fungal species showed $d_N/d_S > 1$ (May et al. 1999; Rau et al. 2007), but the *b1* mating type gene in the mushroom fungus *Coprinus cinereus* was shown to be experiencing balancing selection by comparing the topologies of gene genealogies under balancing selection and neutral scenarios (May et al. 1999). *Het-c* in *Neurospora crassa* was determined to be evolving under balancing selection; evidence included trans-species polymorphisms and an increase in nonsynonymous substitutions in and around the specificity region of *het-c* (Wu et al. 1998). Four codon positions of the WD-40 repeats in *het-d* and *het-e* of *Podospora anserina* have $d_N/d_S > 1$. The authors conclude that balancing selection, rather than directional selection, is operating, because of the high number of amino acid combinations at the four codons of interest (Paoletti et al. 2007).

d_N/d_S ratios > 1 have been found in many plant families, both in gametophytic and sporophytic SI systems (Clark and Kao 1991; Ishimizu et al. 1998; Sato et al. 2002; Takebayashi et al. 2003; Igic et al. 2007; Guo et al. 2011). d_N/d_S ratios > 1 were corroborated with additional data to infer the action of balancing selection: significantly positive Tajima's *D* values, little population structure compared to neutral markers, and low recombination for *SRK* and *SCR* in *Brassica cretica* (Edh et al. 2009), trans-species polymorphisms in *SRK* and *SCR* in several *Arabidopsis* species (Sato et al. 2002; Guo et al. 2011).

MAINTENANCE OF FUHC POLYMORPHISM: NEUTRALITY TESTS

Mean *D* values are negative for both FuHC transcripts regardless of the sampling scheme or how synonymous and nonsynonymous sites are analyzed. A pooling effect resulted in the scattered samples having the lowest values, then pooled samples, then local samples. Arunyawat et al. (2007) predicted this result for genes experiencing a pooling effect, and simulations of the pooling effect gave identical results to those shown here (Stadler et al. 2009).

Negative values of Tajima's *D* are likely due to selection rather than demography, based on two views of the data. First, all of the four sampling schemes suggest that these genes are evolving under a neutral model of evolution, as mean *D* values were consistently close to zero. At the population level, 11 of the 12 housekeeping loci have no populations that were significant for Tajima's *D* when all sites or synonymous sites only were analyzed (few of the housekeeping genes were analyzed for nonsynonymous sites only). The remaining locus (60S ribosomal protein L10) has only a single population with a significant negative value. We detected no pooling effect on the housekeeping genes, so these *D* values should reflect an accurate view of the demographic history of *B. schlosseri* Clade A. Second, the housekeeping loci do not show a consistent negative trend of neutrality tests across all populations, as FuHC does.

Botryllus schlosseri is a Mediterranean endemic that has spread throughout the Atlantic and Pacific oceans via shipping traffic (Lejeusne et al. 2010). The populations sampled in this study are therefore non-native. However, the species does not appear to have experienced a genetic bottleneck, nor does Tajima's *D* provide evidence for recent population growth following a founding event. East and West Coasts of the United States both experienced at least two introductions of *B. schlosseri* (Lejeusne et al. 2010), which could partially explain the apparent lack of a genetic bottleneck.

MAINTENANCE OF FUHC POLYMORPHISM: DISTRIBUTION OF POLYMORPHISM WITHIN AND AMONG DIFFERENT POPULATIONS

The housekeeping genes have substantially more structure among populations within groups and among groups than FuHC. In addition, a majority of the pairwise F_{ST} values are significant for the housekeeping genes, but only one is significant in FuHC.

The population differentiation at the housekeeping loci confirms the significant genetic structure seen at microsatellite loci (Ben-Shlomo et al. 2010; Bock et al. 2012). These results are in sharp contrast to the lack of significant population differentiation at FuHC.

The pattern seen here is consistent with balancing selection acting on FuHC. Loci experiencing balancing selection (which maintains variation) should have larger amounts of polymorphism within populations and smaller amounts among populations than neutral loci (assuming selection pressures are similar between populations), whereas the opposite pattern is expected for loci experiencing directional selection (Schierup et al. 2000).

AMOVA and F_{ST} analyses have been completed in three allorecognition systems in addition to FuHC: *het*/*vic* loci in fungi, SI loci in the Asteraceae (*Guizotia abyssinica*), Brassicaceae (*Arabidopsis* and *Brassica*), and MHC in many vertebrates. In both the chestnut blight fungus (*Cryphonectria parasitica*) and the dry rot fungus (*Serpula lacrymans*), *het*/*vic* loci lacked significant genetic differentiation among different populations (Milgroom and Cortesi 1999; Kauserud et al. 2006).

Patterns at SI loci are similar to those at *het*/*vic* and FuHC. In *G. abyssinica* (niger), 97% of the SI locus variation was found within populations, and F_{ST} values were very low (although statistically significant) (Geleta and Bryngelsson 2010). F_{ST} values are significantly lower when compared to neutral loci, in all cases (*A. lyrata:* Kamau et al. 2007; *A. halleri:* Ruggiero et al. 2008; *B. cretica:* Edh et al. 2009; *B. insularis:* Glemin et al. 2005). These results provide strong evidence for balancing selection driving the evolution of SI loci.

Examinations of polymorphism in the MHC come from birds, fish, and murines. In every study where AMOVAs were conducted, a large percentage of the variation is found within populations (71–96%), and two studies found no evidence for population differentiation (Pfau et al. 2001; Ekblom et al. 2007). The majority of MHC studies report significant population structure at MHC loci (Miller and Withler 1997; Miller et al. 2001; Richman et al. 2003; Campos et al. 2006; Peters and Turner 2008; Koutsogiannouli et al. 2009; Miller et al. 2010).

Researchers have also compared genetic structure between MHC and neutral loci. Studies in chamois, fox, grouse, and rat report lower levels of differentiation for MHC loci than microsatellites, consistent with balancing selection where selection pressures are similar among different populations (Piertney 2003; Sommer 2003; Aguilar et al. 2004; Mona et al. 2008).

MAINTENANCE OF FUHC POLYMORPHISM: CONCLUSIONS

Two sets of hypotheses have generally been advanced for the maintenance of polymorphism at allorecognition loci—those invoking neutral processes, and those invoking selective processes (see Grosberg 1988 for a review). The combination of a selection statistic, neutrality tests, and AMOVA/F_{ST} calculations strongly suggest that FuHC evolves via selection, consistent with studies of many other allorecognition loci. For example, evidence of selection has been shown in the *Hydractinia* locus *alr2* (Rosengarten et al. 2010), *het* loci in fungi (Wu et al. 1998; Paoletti et al. 2007), *lagB1* and *lagC1* in cellular slime molds (Benabentos et al. 2009), and SI loci in several plant families (Sato et al. 2002; Igic et al. 2007).

What is less clear is whether directional or balancing selection is acting on FuHC. We show that FuHC is incredibly polymorphic at both the nucleotide and amino acid level. Unusual levels of polymorphism are taken as evidence of balancing selection (Garrigan and Hedrick 2003). We also show that FuHC has more variation within populations and less variation among different populations than 12 housekeeping loci and less population differentiation than microsatellite loci; these patterns are also consistent with balancing selection (Schierup et al. 2000).

However, the Tajima's *D* values are consistently negative across populations, which is inconsistent with balancing selection (Tajima 1989). Such a pattern could indicate purifying selection, whereby alleles containing deleterious or less fit alleles are kept at a low frequency (Tajima 1989), or a recent selective sweep, whereby strong directional selection causes a particular mutation to fix and reduces the variation in surrounding regions (Maynard-Smith and Haigh 1974).

Exons 1–14 appear to be evolving under directional selection rather than purifying selection. Statistical significance is often the same for *D* values based on synonymous and nonsynonymous sites and no clear pattern emerges when comparing *D* values based on all sites versus synonymous sites. If directional selection leads to a selective sweep, low-frequency variants will appear at both synonymous and nonsynonymous sites, but purifying selection primarily acts on nonsynonymous sites. It is possible that synonymous sites are also experiencing purifying selection. Purifying selection at synonymous sites often acts via codon bias (Ingvarsson et al. 2010). But we see no evidence for codon bias in either Exons 1–14 or Exons 18–29/31, and the two transcripts have very similar values for several codon bias statistics (data not shown).

In contrast, Exons 18–29/31 likely experiences weak purifying selection. *D* values for these exons are significantly negative

only when considering nonsynonymous sites. *D* values based on all sites are consistently less than values based on synonymous sites, which would be expected if purifying selection is acting on nonsynonymous sites. The disparate evolutionary histories of these transcripts can also be seen in Figure S1: Exons 1–14 has substantially higher amino acid diversity than Exons 18–29/31.

How are the contrasting conclusions of AMOVA/F_{ST} and Tajima's *D* to be resolved? Where allorecognition loci evolve by balancing selection, three specific types of selection are possible: spatially variable, heterosis, and negative frequency-dependent (Hedrick 1999). Selection on FuHC is unlikely to vary spatially (in contrast to MHC loci, where parasites may exert different selection pressures on populations). FuHC polymorphism may be maintained by heterosis, or heterozygote advantage. *Botryllus schlosseri* individuals that are heterozygous at FuHC experience a strong fitness advantage relative to homozygotes (De Tomaso 2006). However, this heterosis occurs independently of interindividual interactions: whether FuHC heterozygotes have higher fitness in nature (where fusion is a possibility) is not known.

Fusion introduces a cost to fitness (Rinkevich and Weissman 1992; Chadwick-Furman and Weissman 1995; Chadwick-Furman and Weissman 2003), and individuals heterozygous for FuHC should fuse more than homozygous individuals, thus incurring more cost (De Tomaso et al. 2006). So if FuHC does evolve by balancing selection, negative frequency-dependent balancing selection is the most likely process.

Botryllus schlosseri Clade A is a Mediterranean endemic and has invaded the United States via anthropogenic transport (Lejeusne et al. 2010; Bock et al. 2012). Given the recency of these invasions on an evolutionary timescale, these populations are not likely at equilibrium. An analysis of FuHC sequences from Mediterranean populations could provide further insight into the nature of selection on FuHC.

ACKNOWLEDGMENTS

We thank S. Rendulic for collecting many of the *B. schlosseri* colonies used in this study, and T. McKitrick for advice on FuHC PCR amplification. We also thank A. Cutter and two anonymous reviewers for suggestions on the manuscript. This work was funded by National Science Foundation (NSF) Grant IOS-0842138 to AWDT.

LITERATURE CITED

Aguilar, A., G. Roemer, S. Debenham, M. Binns, D. Garcelon, and R. K. Wayne. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc. Natl. Acad. Sci. USA 101:3490–3494.

Alcaide, M., S. V. Edwards, J. J. Negro, D. Serrano, and J. L. Tella. 2008. Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). Mol. Ecol. 17:2652–2665.

Arunyawat, U., W. Stephan, and T. Stadler. 2007. Using multi-locus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. Mol. Biol. Evol. 24:2310–2322.

Benabentos, R., S. Hirose, R. Sucgang, T. Curk, M. Katoh, E. Ostrowski, J. Strassmann, D. Queller, B. Zupan, G. Shaulsky et al. 2009. Polymorphic members of the *Iag*-gene family mediate kin discrimination in *Dictyostelium*. Curr. Biol. 19:567–572.

Ben-Shlomo, R., E. Reem, J. Douek, and B. Rinkevich. 2010. Population genetics of the invasive ascidian *Botryllus schlosseri* from South American coasts. Mar. Ecol. Prog. Ser. 412:85–92.

Bishop, J. D. D., and A. D. Sommerfeldt. 1999. Not like *Botryllus:* indiscriminate post-metamorphic fusion in a compound ascidian. Proc. R. Soc. Lond. B 266:241–248.

Bock, D. G., H. J. MacIsaac, and M. E. Cristescu. 2012. Multilocus genetic analyses differentiate between widespread and spatially restricted cryptic species in a model ascidian. Proc. R. Soc. Lond. B 279: 2377–2385.

Buss, L. 1982. Somatic cell parasitism and the evolution of somatic tissue compatibility. Proc. Natl. Acad. Sci. USA 79:5337–5341.

Campos, J. L., D. Posada, and P. Moran. 2006. Genetic variation at MHC, mitochondrial, and microsatellite loci in isolated populations of Brown trout (*Salmo trutta*). Conserv. Genet. 7:515–530.

Chadwick-Furman, N. E., and I. L. Weissman. 1995. Life history plasticity in chimeras of the colonial ascidian *Botryllus schlosseri*. Proc. R. Soc. Lond. B 262:157–162.

———. 2003. The effects of allogeneic contact on life-history traits of the colonial ascidian *Botryllus schlosseri* in Monterey Bay. Biol. Bull. 205:133–143.

Clark, A. G., and T.-H. Kao. 1991. Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc. Natl. Acad. Sci. USA 88:9823–9827.

Clark, J. 2003. Plasmodial incompatibility in the myxomycete *Didymium squamulosum*. Mycologia 95:24–26.

Cutter, A. D., G.-X. Wang, H. Ai, and Y. Peng. 2012. Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity. Mol. Ecol. 21:1345–1359.

De Tomaso, A. W. 2006. Allorecognition polymorphism vs. parasitic stem cells. Trends Genet. 22:485–490.

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

Edh, K., B. Widen, and A. Ceplitis. 2009. Molecular population genetics of the *SRK* and *SCR* self-incompatibility genes in the wild plant species *Brassica cretica* (Brassicaceae). Genetics 181:985–995.

Ekblom, R., S. Aresaether, P. Jacobsson, P. Fiske, T. Sahlman, M. Grahn, J. Atlekalas, and J. Hoglund. 2007. Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). Mol. Ecol. 16:1439–1451.

Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Res. 10:564–567.

Garrigan, D., and P. W. Hedrick. 2003. Detecting adaptive molecular polymorphism: lessons from the MHC. Evolution 57:1707–1722.

Geleta, M., and T. Bryngelsson. 2010. Population genetics of self-incompatibility and developing self-compatible genotypes in niger (*Guizotia abyssinica)*. Euphytica 176:417–430.

Geliebter, J., and S. G. Nathenson. 1987. Recombination and the concerted evolution of the murine MHC. Trends Genet. 3:107–112.

Glass, N. L., D. J. Jacobson, and P. K. Shiu. 2000. The genetics of hyphal fusion and vegetative incompatibility in filamentous ascomycete fungi. Annu. Rev. Genet. 34:165–186.

Glemin, S., T. Gaude, M.-L. Guillemin, M. Lourmas, E. Olivieri, and A. Mignot. 2005. Balancing selection in the wild: testing population genetics theory of self-incompatibility in the rare species *Brassica insularis*. Genetics 171:279–289.

Grosberg, R. K. 1988. The evolution of allorecognition specificity in clonal invertebrates. Q. Rev. Biol. 63:377–412.

Grosberg, R. K., D. R. Levitan, and B. B. Cameron. 1996. Evolutionary genetics of allorecognition in the colonial hydroid *Hydractinia symbiolongicarpus*. Evolution 50:2221–2240.

Grosberg, R. K., and J. F. Quinn. 1986. The genetic control and consequences of kin recognition by the larvae of a colonial marine invertebrate. Nature 322:456–459.

Guo, Y.-L., X. Zhao, C. Lanz, and D. Weigel. 2011. Evolution of the *S*-locus region in *Arabidopsis thaliana* relatives. Plant Phys. 157:937–946.

Hedrick, P. W. 1999. Balancing selection and MHC. Genetica 104:207–214.

Hidaka, M., K. Yurugi, S. Sunagawa, and R. A. Kinzie III. 1997. Contact reactions between young colonies of the coral *Pocillopora damicornis*. Coral Reefs 16:13–20.

Hudson, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. 50:245–250.

Hughes, R. N., P. H. Manriquez, S. Morley, S. F. Craig, and J. D. D. Bishop. 2004. Kin or self-recognition? Colonial fusibility of the bryozoan *Celleporella hyalina*. Evol. Dev. 6:431–437.

Igic, B., W. A. Smith, K. A. Robertson, B. A. Schaal, and J. R. Kohn. 2007. Studies of self-incompatibility in wild tomatoes: I. S-allele diversity in *Solanum chilense* Dun. (Solanaceae). Heredity 99:553–561.

Ingvarsson, P. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. Mol. Biol. Evol. 27:650–660.

Ishimizu, T., T. Endo, Y. Yamaguchi-Kabata, K. T. Nakamura, F. Sakiyama, and S. Norioka. 1998. Identification of regions in which positive selection may operate in S-RNase of Rosaceae: implication for *S*-allele-specific recognition sites in S-RNase. FEBS Lett. 440:337–342.

Kamau, E., B. Charlesworth, and D. Charlesworth. 2007. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. Genetics 176:2357–2369.

Kauserud, H., G. P. Saetre, O. Schmidt, C. Decock, and T. Schumacher. 2006. Genetics of self/nonself recognition in *Serpula lacrymans*. Fungal Genet Biol. 43:503–510.

Koutsogiannouli, E. A., K. A. Moutou, T. Sarafidou, C. Stamatis, V. Spyrou, V, and Z. Mamuris. 2009. Major histocompatibility complex variation at class II DQA locus in the brown hare (*Lepus europaeus*). Mol. Ecol. 18:4631–4649.

Lawrence, M. J. 1996. Number of incompatibility alleles in clover and other species. Heredity 76:610–615.

Lejeusne, C., D. G. Bock, T. W. Therriault, H. J. MacIsaac, and M. Cristescu. 2010. Comparative phylogeography of two colonial ascidians reveals contrasting invasion histories in North America. Biol. Invasions 13: 635–650.

Librado, P., and J. Rozas. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

May, G., F. Shaw, H. Badrane, and H. Vekemans. 1999. The signature of balancing selection: fungal mating compatibility gene evolution. Proc. Natl. Acad. Sci. USA 96:9172–9177.

Maynard-Smith, J., and J. Haigh. 1974. The hitchhiking effect of a favorable gene. Genet. Res. 23:23–35.

Mercier, A., Z. Sun, and J.-F. Hamel. 2011. Internal brooding favours premetamorphic chimerism in a non-colonial cnidarian, the sea anemone *Urticina felina*. Proc. R. Soc. Lond. Ser. B 278:3517–3522.

Milgroom, M. G., and P. Cortesi. 1999. Analysis of population structure of the chestnut blight fungus based on vegetative incompatibility genotypes. Proc. Natl. Acad. Sci. USA 96:10518–10523.

Miller, H. C., F. Allendorf, and C. H. Daugherty. 2010. Genetic diversity and differentiation at MHC genes in island populations of tuatara (*Sphenodon* spp.). Mol. Ecol. 19:3894–3908.

Miller, K. M., K. H. Kaukinen, T. D. Beacham, and R. E. Withler. 2001. Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. Genetica 111:237–257.

Miller, K. M., and R. E. Withler. 1997. MHC diversity in Pacific salmon: population structure and trans-species allelism. Hereditas 127: 83–95.

Mona, S., B. Crestanello, S. Bankhead-Dronnet, E. Pecchioli, S. Ingrosso, S. D'Amelio, L. Rossi, P. G. Meneguz, and G. Bertorelle. 2008. Disentangling the effects of recombination, selection, and demography on the genetic variation at a major histocompatibility complex class II gene in the alpine chamois. Mol. Ecol. 17:4053–4067.

Oka, H., and H. Watanabe. 1957. Colony specificity in compound ascidians as tested by fusion experiments. Proc. Japan Acad. Sci. 33:657–664.

Paoletti, M., S. J. Saupe, C. Clave. 2007. Genesis of a fungal non-self recognition repertoire. PLoS One 2:e283.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated-modular approach version 0.4–2. Bioinformatics 26: 419–420.

Paradis, E. 2012. Analysis of phylogenetics and evolution with R. 2nd ed. Springer, New York, USA.

Parham, P., and T. Ohta. 1996. Population biology of antigen presentation by MHC Class I molecules. Science 272:67–74.

Peters, M. B., and T. F. Turner. 2008. Genetic variation of the major histocompatibility complex (MHC class II β gene) in the threatened Gila trout, *Oncorhynchus gilae gilae*. Conserv. Genet. 9:257–270.

Pfau, R. S., R. A. Van Den Bussche, and K. McBee. 2001. Population genetics of the hispid cotton rat (*Sigmodon hispidus*): patterns of genetic diversity at the major histocompatibility complex. Mol. Ecol. 10: 1939–1945.

Piertney, S. B. 2003. Major histocompatibility complex B-LB gene variation in red grouse *Lagopus lagopus scoticus*. Wildl. Biol. 9:251–259.

Rau, D., G. Attene, A. H. D. Brown, L. Nanni, F. J. Maier, V. Balmas, E. Saba, W. Schafer, and R. Papa. 2007. Phylogeny and evolution of mating-type genes from Pyrenophora teres, the causal agent of barley "net blotch" disease. Curr. Genet. 51:377–392.

Richman, A. D., L. G. Herrera, D. Nash, and M. K. Schierup. 2003. Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in *Peromyscus maniculatus*. Genet. Res. 82: 89–99.

Rinkevich, B., R. Porat, and M. Goren. 1995. Allorecognition elements on a urochordate histocompatibility locus indicate unprecedented extensive polymorphism. Proc. R. Soc. Lond. B 259:319–324.

Rinkevich, B., and I. L. Weissman. 1992. Chimeras vs. genetically homogeneous individuals: potential fitness costs and benefits. Oikos 63: 119–124.

Rodi, D. J., S. Mandava, and L. Makowski. 2004. DIVAA: analysis of amino acid diversity in multiple aligned protein sequences. Bioinformatics 20:3481–3489.

Rosengarten, R. D., M. A. Moreno, F. G. Lakkis, L. W. Buss, and S. L. Dellaporta. 2010. Genetic diversity of the allodeterminant *alr2* in *Hydractinia symbiolongicarpus*. Mol. Biol. Evol. 28:933–947.

Ruggiero, M. V., B. Jacquemin, V. Castric, and X. Vekemans. 2008. Hitchhiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. Genet. Res. 90:37–46.

Santelices, B., J. A. Correa, D. Aedo, V. Flores, M. Hormazabal, and P. Sanchez. 1999. Convergent biological processes in coalescing Rhodophyta. J. Phycol. 35:1127–1149.

Sato, K., T. Nishio, R. Kimura, M. Kusaba, T. Suzuki, K. Hatakeyama, D. J. Ockendon, and Y. Satta. 2002. Coevolution of the *S*-locus genes *SRK, SLG*, and *SP11*/*SCR* in *Brassica oleracea* and *B. rapa*. Genetics 162:931–940.

Schierup, M. H., X. Vekemans, and D. Charlesworth. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet. Res. 76:51–62.

Schmidt, G. H. 1982. Aggregation and fusion between conspecifics of a solitary ascidian. Biol. Bull. 162:195–201.

Shaulsky, G., and R. Kessin. 2007. The cold war of the social amoebae. Curr. Biol. 17:684–692.

Sommer, S. 2003. Effects of habitat fragmentation and changes of dispersal behaviour after a recent population decline on the genetic variability of noncoding and coding DNA of a monogamous Malagasy rodent. Mol. Ecol. 12:2845–2851.

Stadler, T., B. Haubold, C. Merino, W. Stephan, and P. Pfaffelhuber. 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182: 205–216.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Takahata, N., Y. Satta, and J. Klein. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. Genetics 130: 925–938.

Takebayashi, N., P. B. Brewer, E. Newbigin, and M. K. Uyenoyama. 2003. Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. 20:1778–1794.

von Salome, J., and J. P. Kukkonen. 2008. Sequence features of *HLA-DRB 1* locus define putative basis for gene conversion and point mutations. BMC Genomics 9:228.

Wilson, D. J., and G. McVean. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172:1411–1425.

Wilson, H. V. 1907. On some phenomena of coalescence and regeneration in sponges. J. Exp. Zool. 5:245–258.

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

Wu, J., S. J. Saupe, and N. L. Glass. 1998. Evidence for balancing selection operating at the *het-c* heterokaryon incompatibility locus in a group of filamentous fungi. Proc. Natl. Acad. Sci. USA 95:12398–12403.

Zeller, R. 2010. Computational analysis of *Ciona intestinalis* operons. Integr. Comp. Biol. 50:75–85.

Associate Editor: A. Cutter

Supporting Information

The following supporting information is available for this article:

Figure S1. Amino acid diversity in FuHC as calculated by the program DIVAA (Rodi et al. 2004) for (a) Exons 1–14 and (b) Exons 18–29/31.

Figure S2. Recombination (ρ) across (a) Exons 1–14 and (b) Exons 18–29/31 of FuHC for the Monterey, CA, population.

Figure S3. Values of the selection statistic from omegaMap 0.5 across (a) Exons 1–14 and (b) Exons 18–29/31 of FuHC for the Monterey, CA, population.

Figure S4. Distribution across FuHC (a) Exons 1–14 and (b) Exons 18–29/31 of the 24 codons that have $> 95\%$ probability of the selection statistic > 1 and are present in at least three of the six populations.

Table S1. Summary statistics for FuHC and 12 housekeeping genes.

Table S2. Values of Tajima's D statistics and associated 95% confidence intervals for all housekeeping genes, asterisks denote statistical significance at $\alpha = 0.05$.

Table S3. AMOVA, fixation indices, and pairwise F_{ST} values for all housekeeping genes.

Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.