Running Head: Balancing selection in Botryllus schlosseri

Balancing selection on allorecognition genes in the colonial ascidian *Botryllus schlosseri*

Marie L. Nydam^{a*}, Emily E. Stephenson^{a,b}, Claire E. Waldman^a, Anthony W. De Tomaso^c

Abstract

Allorecognition is the capability of an organism to recognize its own or related tissues. The colonial ascidian *Botryllus schlosseri*, which comprises five genetically distinct and divergent species (Clades A-E), contains two adjacent genes that control allorecognition: *fuhc^{sec}* and *fuhctm*. These genes have been characterized extensively in Clade A and are highly polymorphic. Using alleles from 10 populations across the range of Clade A, we investigated the type and strength of selection maintaining this variation. Both *fuhc* genes exhibit higher within-population variation and lower population differentiation measures (F_{ST}) than neutral loci. The *fuhc* genes contain a substantial number of codons with > 95% posterior probability of $d_N/d_S > 1$. *fuhc^{sec}* and *fuhctm* also have polymorphisms shared between Clade A and Clade E that were present prior to speciation (trans-species polymorphisms). These results provide robust evidence that the *fuhc* genes are evolving under balancing selection.

Keywords: *Botryllus schlosseri* - allorecognition - balancing selection - *fuhc^{sec}* - *fuhctm* - ascidian

1. Introduction

Allorecognition is the ability of an organism to distinguish self from non-self, or related individuals from non-related individuals (kin discrimination). Well-studied systems of allorecognition are present on all branches of the tree of life: from prokaryotes such as Bacteria (Cao et al., 2015, Rendueles et al., 2015, Stefanic et al., 2015, and Wenren et al., 2013), to unicellular eukaryotes like Archamoebae (Espinosa and Paz-y-Miño-C, 2012), Ciliophora (Cervantes et al., 2013 and Chaine et al., 2010), Dikarya (Brückner

and Mösch, 2012, Goossens et al., 2015, Glass et al., 2000, and Raudaskoski 2015) and Mycetozoa (Kalla et al., 2011, Kaushik et al., 2006, Mehdiabadi et al., 2006, and Romeralo et al. 2012), to multicellular eukaryotes such as Plantae (Allen and Hiscock, 2008, Iwano and Takayama, 2012, Pandey, 1960, and Wu et al., 2013) and Animalia (Flajnik and Kasahara, 2009, Hughes et al., 2004, Karadge et al., 2015, Puill-Stephen et al., 2012, Raftos, 1994, Saito, 2013, Smith et al., 2012, Taketa and De Tomaso, 2015, Westerman et al., 2009). The

genetic basis of allorecognition is not well understood in invertebrates, with the exception of the hydroid *Hydractinia symbiolongicarpus* and the ascidians *Ciona robusta* and *Botryllus schlosseri* (Grosberg and Plachetzki, 2010). In *C. robusta*, allorecognition loci provide a mechanism for inbreeding avoidance (Ban et al., 2005 and Harada et al., 2008), similar to the SI (Self-Incompatibility) loci in angiosperms (Iwano and Takayama, 2012). In *B. schlosseri*, allorecognition loci allow neighboring colonies to determine whether or not they are related; only related colonies fuse to form a new colony.

B. schlosseri comprises five cryptic species: Clades A-E (Bock et al., 2012 and Lopez-Legentil et al., 2006); all previous allorecognition work has focused on Clade A only. Clade A is found in all the major oceans of the world (Bock et al., 2012 and Yund et al., 2015). Clades B-D have only been described from < 10 individuals each, found in 1-2 locations in the western Mediterranean Sea, the southern Bay of Biscay or the English Channel (Bock et al., 2012 and Lopez-Legentil et al., 2006). Clade E is widespread in the northeastern Atlantic Ocean and Mediterranean Sea (Bock et al., 2012 and Lopez-Legentil et al., 2006). The species within the *B. schlosseri* complex are genetically divergent:

10.8-16.5% at the mitochondrial cytochrome oxidase I locus (mtCOI; Bock et al., 2012).

Bayesian clustering of seven microsatellite loci revealed no evidence of gene flow between Clades A, D or E (Clades B and C were not analyzed) (Bock et al., 2012).

The Clade A

allorecognition determinant was initially thought to be a single gene with two alternative transcripts (De Tomaso et al., 2005) but has since been characterized as two genes separated by 227 base pairs (Nydam et al., 2013a). *fuhc^{sec}* encodes a 555 amino acid protein that is secreted from the cell, and *fuhctm* encodes a 523 amino acid membrane-bound protein (Nydam et al., 2013a). Both proteins are highly polymorphic, although *fuhc^{sec}* is more variable than *fuhctm* (Nydam et al., 2013b). Polymorphisms in both proteins predict allorecognition outcome. However, *fuhc^{sec}* may be determining or driving histocompatibility (Nydam et al., 2013a). Evidence for this possibility comes from repeated pairings between two genotypes that result in rejections so weak that they are often mistaken as non-reactions (Nydam et al., 2013a). Rejection severity is thought to be linked to divergence between alleles: more amino acid differences between alleles resulting in stronger rejections (Scofield and Nagashima, 1983). In this weak rejection, the *fuhc^{sec}* alleles were nearly identical but the *fuhctm* alleles were divergent.

The

extensive polymorphism in allorecognition loci is maintained by natural selection, either balancing or directional (reviewed in Nydam and De Tomaso, 2011). Balancing selection maintains polymorphism in a population by pushing an allele to an intermediate frequency, whereas directional selection reduces polymorphism in a population by moving an allele to fixation (Hedrick, 2007). Several population genetic tests are

commonly used to detect selection; in some cases these tests can also differentiate between balancing and directional selection. A summary of the tests used in the allorecognition literature and in this study is presented in Table 1.

In several allorecognition systems, balancing selection has maintained polymorphisms that were present prior to speciation (Figueroa et al., 1988, Guo et al., 2011, Ioerger et al., 1990, Sato et al., 2002, and Wu et al., 1998). Selection has been shown to operate on Clade A *fuhc^{sec}* and *fuhctm* when sequences from the eastern Pacific and northwestern Atlantic Oceans were analyzed. ω statistics (d_N/d_S) revealed codons throughout *fuhc^{sec}* and *fuhctm* with > 95% probability of $\omega > 1$ and values of Tajima's D that deviated from neutrality at allorecognition loci but not at any of 12 housekeeping loci (Nydam et al., 2013b). Additionally, a statistically significant lack of population structure at both *fuhc^{sec}* and *fuhctm* was contrasted with strong structure at 12 housekeeping loci (Nydam et al., 2013b), mtCOI (Lopez-Legentil et al., 2006), and microsatellites (Ben-Shlomo et al., 2010 and Bock et al., 2012). However, the type of selection could not be identified with confidence: ω statistics cannot differentiate between balancing and directional selection, Tajima's D statistics provided evidence for directional selection, and analyses of population structure pointed towards balancing selection (Nydam et al., 2013b).

The current study aims to conclusively determine whether balancing or directional selection is acting on Clade A *B. schlosseri* allorecognition loci. In our previous study, we only included Clade A populations from the eastern Pacific and northwestern Atlantic Oceans. This limited sampling could explain the inconsistent patterns of selection inferred in Nydam et al.

2013b. Here, we include Clade A sequences from four new populations (two from the English coast of the English Channel and two from the French coast of the English Channel). We can now analyze selection patterns in 10 populations, which may result in a stronger consensus about which type of selection is most prevalent in the species as a whole. Conversely, populations in different geographic regions may be experiencing different types of selection. Adding a new geographic region allows us to contrast the evolutionary history of Clade A *fuhc^{sec}* and *fuhctm* in three geographic regions (eastern Pacific, northwestern Atlantic, and English Channel). Inclusion of only Clade A sequences in Nydam et al., 2013b limited the scope of the population genetic analyses we could perform. In the current study, we include Clade E *fuhc^{sec}* and *fuhctm* sequences to make inferences of selection that were not possible with intraspecific data from Clade A, including tests of trans-species polymorphism.

2. Materials and methods

2.1. Sampling

Clade A and Clade E colonies were collected from floating docks in the eastern Pacific, northwestern Atlantic, Bay of Biscay, Mediterranean, and English Channel from 2009-2015 (Table 2). Colonies from Clades B, C and D were not recovered, despite collecting in all locations where they had been previously found. Colonies were collected from several locations in each harbor or marina in order to include a broad representation of the *fuhc* alleles in the population. Single systems were cut from individual colonies and immediately placed in RNeasy® (Thermo Fisher Scientific, Waltham, MA, USA). Samples were stored at -80°C within 12 days of collection.

2.2. Amplification and sequencing: *fuhc* genes

Because morphological characters differentiating Clade A and Clade E have not been identified (Bock et al. 2012), all samples were first sorted into clades by examining the sequences of two genes: mitochondrial cytochrome oxidase I (mtCOI) and nuclear 18S rRNA. To amplify these two genes, genomic DNA was extracted from whole systems using a Nucleospin Tissue Kit® (Macherey-Nagel, Düren, Germany). The mtCOI gene was amplified using the universal invertebrate primers LCO1490 and HCO2198 (Folmer et al., 1994). The 18S rRNA gene was amplified using 557F and 3'R (Teske et al., 2011). PCR amplification was performed in a GenePro Thermal Cycler (Bulldog Bio, Portsmouth, NH, USA) using a 10- μ l total reaction volume with 2 mM MgCl₂, 0.2 mM dNTPs, 1 μ l of 10x buffer (50mM KCl, 20mM Tris (pH 8.4)), 0.2 μ M of each primer, 0.08 U of Taq Polymerase (New England Biolabs, Ipswich, MA) and 1 μ l of template DNA. The mtCOI PCR protocol was as follows: 35x (95°C for 30 seconds, 51°C for 1 minute, 72°C for 1 minute). The 18S PCR protocol was as follows: 35x (95°C for 30 seconds, 54°C for 1 minute, 72°C for 2 minutes). PCR products were incubated with 0.5 μ l each of Exonuclease I (New England Biolabs, Ipswich, MA) and Antarctic Phosphatase (New England Biolabs, Ipswich, MA) at 37°C for 45 min, followed by 90°C for 10 min. The PCR products were sequenced at the University of California Berkeley Sequencing Facility and the University of Kentucky's Advanced Genetic Technologies Center using ABI-3730 automated sequencers (Applied Biosystems, Foster City, CA).

Once each individual was identified as Clade A or Clade E, total RNA was extracted from frozen tissue using the NucleoSpin Nucleic Acid and Protein Purification Kit (Macherey-Nagel, Düren,

Germany). This RNA was used to synthesize single-stranded cDNA using SuperScript III reverse transcriptase (Invitrogen, Carlsbad, CA, USA) and an oligo (dT) primer. PCR primers for *fuhc^{sec}* are WholelgF2: 5' GAAATGTTGCTGAAAATATTCTGTCTT 3' and 14-18soeR: 5' TACTTCAAGTCGACAGTTCCAATCAACGTA 3'. PCR primers for *fuhctm* are 14-18soeF: 5' TACGTTGATTGGAAGTGTGCGACTTGAAGTA 3' and WholelgR2: 5' GATACTTGGCTCTCGCCTTGATCTT 3'. The *fuhc^{sec}* PCR protocol was 35x (95°C for 30 seconds, 51°C for 30 seconds, 72°C for 2 minutes), 72°C for 5 minutes, and the *fuhctm* PCR protocol was 35x (95°C for 30 seconds, 55°C for 30 seconds, 72°C for 2 minutes), 72°C for 5 minutes. 20 µl PCRs contained 13.6 µl of H₂O, 4 µl of 5x HF buffer (Thermo Fisher Scientific, Waltham, MA, USA), 0.2mM dNTPs, 0.6 µl of 100% DMSO, 0.3333 µM of each primer, 0.02U/µl of Phusion Polymerase (Thermo Fisher Scientific, Waltham, MA, USA), and 2 µl of template DNA. PCR products were cloned using the pGEM®T kit (Promega, Madison, WI).

48 clones from each cloning reaction were used as template for PCR amplifications using m13 primers; the length of the insert in each clone was then visualized on an agarose gel. All inserts of the correct length were sequenced, unless > 12 of the 48 clones contained inserts of the correct length. In these cases, 12 clones containing inserts of the correct length were randomly picked for sequencing. Inserts were sequenced in both the forward and reverse directions using m13 primers. Forward and reverse sequences for each clone were combined and manually edited to create a consensus sequence. All inserts for each individual were aligned, and alleles were identified from the alignment. Sequences of alleles from each individual were submitted to GenBank: *fuhc^{sec}* JN082923-JN083035, KX649238-KX649339, *fuhctm*

JN082923-JN083035, KX649340-KX649462. Sequences were aligned using Muscle (MUltiple Sequences Comparison by Log-Expectation) as implemented in the Aligner 6.0.2 software (Codon Code Corporation, Centerville, MA). Alignments were then manually edited.

2.3. Amplification and sequencing: housekeeping genes

For each eastern Pacific and northwestern Atlantic population, the following 12 housekeeping genes were amplified: 40S ribosomal protein 3A, 60S ribosomal protein L6, 60S ribosomal protein L8, 60S ribosomal protein L10, 60S ribosomal protein L13, ADP/ATP translocase 3, adult-type muscle actin 2, cytoplasmic actin 2, heat shock cognate 71 kDa protein, heat shock protein HSP-90 beta, vasa, and vigilin. These genes were chosen because they are not expected to evolve via natural selection: any deviation from neutral evolution in these genes would signify demographic events that have affected the entire genome rather than selection at individual loci.

For each English Channel population, four housekeeping genes (60S ribosomal protein L6, 60S ribosomal protein L13, adult-type muscle actin 2, heat shock cognate 71 kDa protein) were amplified. These four housekeeping genes were chosen because they were representative of the overall pattern of population structure exhibited by the 12 housekeeping genes sequenced for populations in the northwestern Atlantic and eastern Pacific regions (Nydam et al. 2013b). In the northwestern Atlantic and eastern Pacific analysis, the four housekeeping genes had an average 71% of variation within populations, compared to an average 73% of variation within populations in the 12 housekeeping genes (Nydam et al., 2013b).

PCR protocols for all housekeeping genes are listed in Supplementary Table 1. PCR products were sequenced at the University of California Berkeley Sequencing Facility and the University of Kentucky's Advanced Genetic Technologies Center using ABI-3730 automated sequencers (Applied Biosystems, Foster City, CA). All sequences were phased using DnaSP 5.10.1 (Librado and Rozas, 2009). All sequences have been submitted to GenBank (40S ribosomal protein 3A: JQ596880-JQ596936, 60S ribosomal protein L6: JQ596937-JQ597084, KX649463 - KX649549, 60S ribosomal protein L8: JQ597085-JQ597174, 60S ribosomal protein L10: JQ597175-JQ597294, 60S ribosomal protein L13: JQ597595-JQ597716, KX649550 - KX649631, heat shock cognate 71 kDa protein: JQ597295-JQ597430, KX649632 - KX649701, 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, KX649702 - KX649761, vasa: JN083304-JN083376, vigilin: JQ597975-JQ598070). Sequences were aligned and edited following the *fuhc* methods above.

2.4. Analyses

2.4.1. Selection

The following Clade A populations from Table 2 were used for all selection analyses in both *fuhc* genes: eastern Pacific populations (Monterey CA USA, Santa Barbara CA USA and Seattle WA USA), northwestern Atlantic populations (Falmouth MA USA, Quissett MA USA, Sandwich MA USA), English Channel populations (Brest France, Concarneau France, Hamble England, Poole England). The polymorphism statistics π (nucleotide diversity) and number of haplotypes for each Clade A population for all

genes in this study were calculated in DnaSP 5.10.1 (Librado and Rozas, 2009) (Supplementary Table 2).

Both Fu and Li's and Tajima's neutrality statistics were used to assess selection on *fuhc^{sec}* and *fuhctm* in Clade A populations (Fu and Li, 1993 and Tajima 1989). These statistics were calculated using DnaSP 5.10.1 (Librado and Rozas, 2009). Negative values of Fu and Li and Tajima's statistics are evidence for directional or purifying selection, and positive values for balancing selection (Fu and Li, 1993, Tajima, 1989).

In highly polymorphic loci such as *fuhc^{sec}* and *fuhctm*, elevated mutation rates can cause positive values of neutrality test statistics independent of balancing selection (Cutter et al., 2012 and Stadler et al., 2009). This effect can be recognized 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).

To test for a pooling effect, all neutrality test statistics were calculated for each population individually (local), and for populations pooled in three different ways: 1) all populations in the entire species pooled, 2) all populations in each region (eastern Pacific, northwestern Atlantic, English Channel) pooled, 3) English populations in the English Channel pooled, and the French populations in the English Channel pooled. Cutter et al. (2012) provided a model for the pooling schemes. Each of the four analyses (one local and three pooled) was run three times: with non-synonymous and synonymous sites, non-synonymous sites only, and synonymous sites only. P values for each analysis were obtained through 1,000 coalescent simulations using θ , because

preliminary analyses showed that simulations using θ and segregating sites gave similar results. $\theta = 4N\mu$, where N is the effective population size, and μ is the mutation rate. Per gene recombination (R) for each population was determined in DnaSP 5.10.1 (Librado and Rozas, 2009) and used in the coalescent simulations.

We were not able to amplify *fuhc^{sec}* from other botryllid species to use as an outgroup for Fu and Li's D and F statistics. We therefore calculated D^* and F^* , which do not use outgroup sequences. For *fuhctm*, we used a sequence from *Botrylloides diegensis* as an outgroup for Fu and Li's D and F statistics.

The Tajima's D neutrality statistic was also calculated for each housekeeping gene sequenced in this study using DnaSP 5.10.1 (Librado and Rozas, 2009). P values were obtained as described above for the *fuhc* loci. None of the 12 housekeeping genes were found to show a pooling effect in the northwestern Atlantic or eastern Pacific regions (Nydam et al., 2013b), so neutrality test were conducted on local samples only.

The McDonald Kreitman test (McDonald and Kreitman, 1991) was used to compare the synonymous substitutions within and between Clades A and E to the nonsynonymous substitutions within and between Clades A and E, to test for selection operating on *fuhc^{sec}* and *fuhctm*. The MK test was executed in DnaSP 5.10.1 (Librado and Rozas, 2009). The HKA (Hudson, Kreitman and Aguade) test was used to compare polymorphism and divergence in *fuhc^{sec}* and *fuhctm* to polymorphism and divergence at a neutral locus: 60S ribosomal protein L6 (Hudson et al., 1987). The HKA test direct mode option was selected in DnaSP 5.10.1.

Using the program omegaMap 0.5 (Wilson and McVean, 2006), the selection statistic ω (d_N/d_S) and 95%

HPD (high density probability) regions were calculated across *fuhc^{sec}* and *fuhctm* for all 10 Clade A populations. $\omega > 1$ indicates balancing or directional selection, whereas $\omega < 1$ signals purifying selection. The posterior probability of selection per codon was also determined for each gene. The analyses were run by the Computational Biology Service Unit at Cornell University. Each run comprised 250,000 iterations, with every 1000th iteration recorded. The following distributions were used: improper inverse for μ (rate of synonymous transversion) and κ (transition–transversion ratio), inverse for ω (selection parameter) and p (recombination rate). Initial parameter values were 0.1 for μ and 3.0 for κ . ω and p priors ranged from 0.01 to 100. ω values could vary across sites, through the use of an independent model. The traces of μ and κ were plotted to decide the range of burn-in for each run. Two runs were produced for each population for each gene. These two runs were always combined, because the mean and 95% HPD regions for each parameter in the two runs were similar in every case. Exons 5 and 6 for *fuhc^{sec}* and Exons 3 and 10 for *fuhctm* were determined to have at least two codons that had $> 95\%$ posterior probability of $\omega > 1$ in a majority of the populations. Selection statistic values for these exons as well as exon regions containing predicted domains (*fuhc^{sec}*: Exons 8-9, *fuhctm*: Exons 1-3, 4-5, 10) were compared with selection statistic values averaged across the rest of the gene using a one-tailed Mann–Whitney *U*-tests in R 3.3.1 (R Core Team 2016).

Assuming selection pressures are similar between populations, populations experiencing balancing selection at *fuhc^{sec}* and *fuhctm* should have higher variation within populations and lower differentiation between populations than neutral loci (and lower variation within populations and high differentiation between populations for

directional selection) (Schierup et al., 2000). Balancing selection maintains variation within a population, as opposed to purifying or directional selection, which results in decreased variation within a population.

Distribution of polymorphism within and among Clade A populations was investigated for *fuhc^{sec}*, *fuhctm*, and all amplified housekeeping genes in Arlequin 3.5.2.2 (Excoffier and Lischer, 2010). AMOVA (Analysis of MOlecular VAriance), fixation indices (F_{CT} , F_{SC} , F_{ST}) and population pairwise F_{ST} values were determined. Each region (eastern Pacific, northwestern Atlantic, English Channel) was analyzed separately. The eastern Pacific and northwestern Atlantic populations were each combined into a single group, so variation was apportioned within and among populations only. The English Channel populations were separated into English (Hamble and Poole) and French (Brest and Concarneau) groups, so variation was split among groups, among populations within groups, and within populations. Because the comparison of population structure between *fuhc* genes and neutral loci allows inference of selection (Schierup et al., 2000), AMOVA, fixation indices and population pairwise F_{ST} values were calculated for all amplified housekeeping genes.

Population structure was also analyzed using STRUCTURE v.2.3.4 (Hubisz et al., 2009). Two data files were constructed: one for the 12 Clade A nuclear housekeeping genes plus mitochondrial COI, and one for the Clade A *fuhc* genes. Sequence data was recoded as follows: for each gene, each unique haplotype was given a numerical code (i.e. 1:n, where n is the number of unique haplotypes sequenced for a particular gene) (Pritchard, 2010). Each individual's haplotype was then determined, and the numerical code of the individual's haplotype was then

recorded as an allele in the data file. An admixture ancestry model was employed.

Admixture linkage disequilibrium was not modeled and prior population information (i.e. LOCPRIOR and USEPOPINFO) was not used. Allele frequencies were assumed to be independent, with $\lambda = 1$.

Analyses

were run for $K = 1:10$, because there are 10 Clade A populations in this study. The appropriate value of K was determined by comparing the estimated log probability of the data for each K . For the housekeeping genes data file, $K = 3$ had the largest estimated log probability (-3540.9) and no other K values had log probabilities within 5-10 points of the log probability for $K = 3$. In this case, $K = 3$ is the appropriate value of K (Pritchard, 2010). For the *fuhc* genes data file, estimated log probabilities for $K = 1:6$ were very similar (between -1620 and -1636), and were all substantially larger than the estimated log probabilities for $K = 7:10$ (between -1686 and -1875). In this case, the smallest value of K should be chosen ($K = 1$) (Pritchard, 2010). We also examined the values of Q for $K = 2$, because $K = 2$ had the largest estimated log probability. The estimated log probability of $K = 1$ was -1626.7, and of $K = 2$ was -1620.5.

Burn-in was 10,000 steps, with a total run length of 100,000 steps. Summary statistics of α and likelihood parameters were assessed to determine the appropriate burn-in and run length. Q , which is the proportion of each individual's ancestry attributed to population K , was graphed using STRUCTURE PLOT v.2.0 (Ramasamy et al., 2014).

2.4.2. Trans-species Polymorphism

Several analyses were conducted to investigate whether trans-species polymorphism (retention of polymorphism present in the ancestor of Clade A and Clade E) exists in this system. First, the number of fixed differences between Clade A and Clade E in *fuhc^{sec}* and *fuhctm* was calculated and compared to the number of fixed differences between Clade A and Clade E in mtCOI. Lack of fixed differences suggests that mutations are maintained at intermediate frequencies by balancing selection (Garrigan and Hedrick, 2003). Second, d_S (average number of synonymous substitutions per synonymous site) was calculated for Clade A *fuhc^{sec}* and *fuhctm* alleles and compared to d_S for Clade A and Clade E in four housekeeping genes (60S ribosomal protein L6, 60S ribosomal protein L13, adult-type muscle actin 2, heat shock cognate 71 kDa protein). d_S values were calculated in DnaSP 5.10.1 (Librado and Rozas, 2009). This analysis is based on the prediction that intraspecific d_S values from *fuhc* genes will be larger than interspecific d_S values from neutral loci, if *fuhc* alleles pre-date species (Ioerger et al., 1990).

Third, Maximum Likelihood trees of Clade A and Clade E individuals based on *fuhc^{sec}*, *fuhctm*, 18S rRNA and mtCOI were created in MEGA 6.0 (Tamura et al., 2013) in order to visualize the extent to which alleles have sorted in the four genes. Recent introgression or ancestral polymorphism maintained by balancing selection in the *fuhc* genes can be inferred from a pattern of incomplete lineage sorting in *fuhc* genealogies as compared to the 18S rRNA or mtCOI genealogies. A GTR+I+G (General Time Reversible, Gamma distributed with invariant sites) model was used. 1,000 bootstrap replicates were obtained.

3. Results

3.1. Selection

For Fu and Li's D^* and F^* and Tajima's D neutrality tests of $fuhc^{sec}$, a consistent pooling effect was detected for the eastern Pacific region, the northwestern Atlantic region, and the English populations in the English Channel region. Local values were consistently higher than pooled values when all sites, nonsynonymous sites only, and synonymous sites only were considered (Supplementary Table 3). Therefore, pooled values were used (Table 3) because local values are skewed towards an artefactual excess of intermediate frequency alleles. Of the three pooling schemes presented in the Methods, we chose the third scheme: all populations in the eastern Pacific pooled, all populations in the northwestern Atlantic pooled, English populations in the English Channel pooled, and the French populations in the English Channel pooled. The first scheme (all Clade A populations pooled) does not allow us to examine differences between regions. The second scheme (all populations in each of the three regions pooled) pools all English Channel populations together, and we determined that the English and French populations have distinct values for neutrality test statistics (Supplementary Table 3).

For Fu and Li's D and F neutrality tests of $fuhc^{tm}$, a consistent pooling effect was detected in the eastern Pacific region and the northwestern Atlantic region (Supplementary Table 3). A consistent pooling effect was detected for all three regions using Tajima's D (Supplementary Table 3). Therefore, pooled values were used (Table 3) because local values are skewed towards an artefactual excess of intermediate frequency alleles. We chose the third pooling scheme for the same reasons presented for $fuhc^{sec}$ above.

The eastern Pacific and northwestern Atlantic regions have negative $fuhc$ values for all

three statistics in nearly every analysis. The results for the English Channel are more variable. For *fu_hc^{sec}*, English values are negative and French values are positive. For *fu_hctm*, English values are positive for Fu and Li's statistics but negative for Tajima's D. French values are negative for nonsynonymous sites, and positive for all sites and synonymous sites.

Based on neutrality test statistics, selection on both *fu_hc* genes, if present, is weak in the English Channel. None of the neutrality test values are statistically different from zero, regardless of which neutrality test is employed or whether all sites, nonsynonymous sites or synonymous sites are considered. This is in sharp contrast to the northwestern Atlantic populations, where all *fu_hc^{sec}* neutrality tests are statistically significant regardless of which sites are considered, and where *fu_hctm* tests are statistically significant for all sites and nonsynonymous sites. The English Channel results are also in contrast to the eastern Pacific region, where Fu and Li's D* for *fu_hc^{sec}* using nonsynonymous sites and Fu and Li's D and F using all sites and Tajima's D using nonsynonymous sites are different from zero at $\alpha = 0.05$.

In neutrality tests, deviations from zero can indicate demographic effects as well as selection. However, demographic events like rapid population expansion after a bottleneck or subdivided populations would be detected genome-wide. No housekeeping genes showed any statistical deviation from zero for Tajima's D in any of the populations used in this study, except for one gene in one population (60S ribosomal protein L10 in Sandwich, MA, USA; Supplementary Table 4).

There were no fixed differences (either synonymous or nonsynonymous) between Clade A and Clade E in either *fu_hc^{sec}* or *fu_hctm*. As the McDonald Kreitman test assesses heterogeneity in a 2x2 matrix, the

test cannot be run if two of the four cells in the matrix are zero. Neither the HKA test for *fuhc^{sec}* nor *fuhctm* provided evidence of non-neutral evolution. The HKA X^2 value for *fuhc^{sec}* was 0.173 (p value = 0.68) and for *fuhctm* was 0.98 (p value = 0.32).

ω statistics for both *fuhc^{sec}* and *fuhctm* are shown in Table 4. The average number of *fuhc^{sec}* codons with > 95% posterior probability of $\omega > 1$ is 17 for the English Channel populations, 25 for the northwestern Atlantic, and 28 for the eastern Pacific. The codons in Exon 5 have a statistically higher ω value than the rest of the transcript for Hamble, England in the English Channel, all northwestern Atlantic populations, and Seattle, WA, USA in the eastern Pacific. The codons in Exon 6 have a statistically higher ω value than the rest of the transcript for Falmouth, MA, USA and Sandwich, MA, USA in the northwestern Atlantic. ω values in Exons 8 and 9 do not differ from the rest of the transcript in any populations.

The average number of *fuhctm* codons with > 95% posterior probability of $\omega > 1$ is 12 for the English Channel populations, 16 for the northwestern Atlantic, and 13 for the eastern Pacific. The codons in Exon 3 have a statistically higher ω value than the rest of the transcript for Quissett, MA, USA in the northwestern Atlantic. The codons in Exon 10 have a statistically higher ω value than the rest of the transcript for Hamble, England in the English Channel. The codons in Exon 4-5 have a statistically higher ω value than the rest of the transcript for Seattle, WA, USA in the eastern Pacific. ω values in Exons 1-3 analyzed together and Exons 1-5 analyzed together do not differ from the rest of the transcript in any populations.

AMOVA-based population structure analyses clearly show that all three regions have higher within population variation in their *fuhc* genes than in housekeeping genes averaged together (Table 5).

By region, the northwestern Atlantic shows that *fuhc^{sec}* has higher within population variation than 8/12 housekeeping genes and *fuhctm* has higher within population variation than 7/12 housekeeping genes (Supplementary Tables 5 and 6). In the eastern Pacific populations, *fuhc^{sec}* has higher within population variation than 12/12 housekeeping genes and *fuhctm* has higher within population variation than 8/12 housekeeping genes (Supplementary Tables 7 and 8). *fuhc^{sec}* and *fuhctm* both have higher within population variation than 3/4 housekeeping genes in the English Channel (Supplementary Tables 9 and 10).

The F_{ST} and population pairwise F_{ST} values corroborate the lack of population structure in the *fuhc* genes compared with the neutral loci: in every region these values are always lower in the *fuhc* genes than in the housekeeping genes averaged (Table 5). By region, the northwestern Atlantic has no significant fixation indices for either *fuhc* gene, with the exception of F_{ST} in *fuhctm* (Supplementary Table 5). There are no significant population pairwise F_{ST} values for *fuhc^{sec}*, and 1/3 for *fuhctm* (Supplementary Table 5). In contrast, 7/12 housekeeping genes have significant F_{ST} values, 2 genes have 1/3 significant population pairwise F_{ST} values, and 5 genes have 2/3 significant population pairwise F_{ST} values (Supplementary Table 6). In eastern Pacific populations, there are no significant fixation indices for either *fuhc* gene, with the exception of F_{ST} in *fuhctm* (Supplementary Table 7). There are no significant population pairwise F_{ST} values for *fuhc^{sec}*, and 2/3 for *fuhctm* (Supplementary Table 7). In contrast, 10/12 housekeeping genes have significant F_{ST} values, 4 genes have 1/3 significant population pairwise F_{ST} values, 4 genes have 2/3 significant population pairwise F_{ST} values, and 2 genes have 3/3 significant population pairwise F_{ST} values (Supplementary Table 8). In the English Channel, there are no

significant fixation indices or population pairwise F_{ST} values for either *fuhc* gene (Supplementary Table 9). In contrast, 3/4 housekeeping genes have significant values of fixation indices and 3/4 genes have at least 1/3 significant population pairwise F_{ST} values (Supplementary Table 10).

Model-based clustering analyses of population structure employed in STRUCTURE provide further unambiguous evidence that *fuhc* genes have lower population structure than housekeeping genes (Fig. 1). The housekeeping genes have considerable population structure (Fig. 1a). The K with the highest likelihood was 3, which corresponds to the three regions used in the AMOVA analyses: English Channel (Populations 1-4 on STRUCTURE plot), northwestern Atlantic (Populations 5-7), and eastern Pacific (Populations 8-10). In contrast to the housekeeping genes, the *fuhc* genes have no discernable population structure (Fig. 1b, c). When $K = 1$, all individuals come from a single population. When $K = 2$, all individuals are assigned ~0.5 of their ancestry to one population, and ~0.5 to the second population. When there is no population structure, the proportion assigned to each individual is roughly equal (e.g. 50% to each population if $K = 2$ and 33% to each population if $K = 3$) and most individuals will display admixture (Pritchard, 2010).

3.2. Trans-species Polymorphism

In the first analysis of trans-species polymorphism, as revealed in the McDonald-Kreitman test, neither *fuhc* gene has a single fixed difference between Clade A and Clade E. The *fuhc^{sec}* alignment is 991 bp long and the *fuhctm* alignment is 881 bp long.

This is in contrast to the mtCOI gene, which contains 39 fixed differences between Clade A and E. The mtCOI alignment is 577 bp long. Second, the average d_s value for Clade A *fuhc^{sec}* is 0.073 and for Clade A *fuhctm* is 0.145. The d_s value for Clade A and Clade E for four housekeeping genes averaged is 0.010. The synonymous substitutions are much higher within Clade A in *fuhc* genes than between Clade A and Clade E in neutrally evolving genes.

Third, the patterns of lineage sorting are distinct in *fuhc^{sec}*, *fuhctm*, and 18S rRNA genealogies. Clade A and Clade E individuals are not well-sorted when *fuhc^{sec}* or *fuhctm* alleles are used to build the genealogies (Figs. 2 and 3), but are when 18S rRNA is used (Fig. 4). Clade A and Clade E are not reciprocally monophyletic when *fuhc^{sec}* or *fuhctm* alleles are used to build the phylogeny (Figs. 1 and 2), but are when mtCOI is used (Fig. 5). Some phylogenetic structure can be seen in the *fuhctm* tree, in the form of a well-supported clade that contains 2 Clade A alleles and 31 Clade E alleles (Fig. 2). In contrast, there are no *fuhc^{sec}* clades composed primarily of Clade E alleles (Fig. 3).

4. Discussion

Evidence for selection on *fuhc^{sec}* in each region and then *fuhctm* in each region will be considered here; Table 6 is a summary of the selection inferences on each region. ω provides evidence for strong balancing or directional selection acting on *fuhc^{sec}* in the northwestern Atlantic region. The ω statistic is not designed to differentiate between balancing and directional selection when $\omega > 1$, however. AMOVA, fixation indices, and

pairwise population F_{ST} values show that variation within populations at *fuhc^{sec}* is elevated with respect to neutrally evolving housekeeping genes. Model-based clustering analyses also display a complete lack of population structure at *fuhc* genes, in contrast to housekeeping genes in which the northwestern Atlantic individuals derive primarily from northwestern Atlantic lineages (Fig. 1). These analyses provide evidence for balancing selection (Schierup et al., 2000). On the other hand, neutrality statistics provide evidence for directional selection. Fu and Li's D^*/F^* and Tajima's D statistics have similar negative values whether all sites, nonsynonymous sites, or synonymous sites are evaluated. After a selective sweep caused by directional selection, mutations will occur in both synonymous and nonsynonymous sites. In this case, the values of neutrality statistics will be negative (indicating low frequency mutations) for both synonymous and nonsynonymous sites, and will not vary substantially between the two types of sites. Purifying selection can also result in negative values of neutrality statistics for both synonymous and nonsynonymous sites. Purifying selection on synonymous sites acts through codon bias, and we could find no evidence for codon bias in either *fuhc* gene (Nydam et al., 2013b). Indication of strong balancing or directional selection acting on *fuhc^{sec}* in eastern Pacific populations comes from ω . However, only Exon 5 in the Seattle population has an elevated ω value. Distribution of polymorphism within and among populations supports balancing selection, as does the lack of population structure at *fuhc^{sec}* compared to the strong population structure in the Monterey and Santa Barbara populations (Fig. 1). Neutrality statistics provide evidence for purifying selection acting on *fuhc^{sec}* in the eastern Pacific. Fu and Li's D^*/F^* statistics have more negative values when nonsynonymous sites are

used than when synonymous sites are used. This indicates more low frequency variants at nonsynonymous sites: purifying selection primarily acts on nonsynonymous sites. Larger negative values at synonymous sites for Tajima's D suggests purifying selection on synonymous sites, but no codon bias was detected in eastern Pacific populations at *fuhc^{sec}* (Nydham et al., 2013b). The ω statistic

identifies codons under balancing or directional selection in the English Channel populations for *fuhc^{sec}*. Selection strength is moderate, based on the number of codons with > 95% probability of $\omega > 1$. Also, only Exon 5 in the Hamble population has a ω value that is statistically higher than the other portions of the transcript. AMOVA, fixation indices, pairwise population F_{ST} values and clustering analyses display a complete lack of population structure at *fuhc*, as expected if balancing selection is operating. English populations have values very close to zero, and no departure from neutrality can be inferred. In contrast, French populations have values > 0, which indicates an excessive intermediate frequency mutations caused by balancing selection.

When comparing the evolutionary history of the three regions at *fuhc^{sec}*, the ω statistic reveals stronger balancing or directional selection across the gene for eastern Pacific and northwestern Atlantic populations than English Channel populations, and stronger selection acting on Exons 5 and 6 in the northwestern Atlantic region. There are fewer codons with > 95% posterior probability of selection ($\omega > 1$) in the English Channel populations than in the northwestern Atlantic populations. Additionally, Exon 5 has statistically higher ω values than the rest of *fuhc^{sec}* in 100% of the northwestern Atlantic populations, but in only 25% of the English Channel populations and 33% of the eastern Pacific populations. The

average W value is also much higher in the northwestern Atlantic region than in either the eastern Pacific or English Channel regions. 66% of the northwestern Atlantic populations have Exon 6 ω values that are higher than the rest of the gene, but Exon 6 ω values do not stand out in either the eastern Pacific or English Channel populations.

Fu and Li's D^*/F^* and Tajima's D also uncover stronger selection in the eastern Pacific and northwestern Atlantic populations than the English Channel populations. According to these neutrality statistics, *fuhc^{sec}* appears to be evolving neutrally in English populations, via balancing selection in French populations (positive values of neutrality test statistics), via directional selection in northwestern Atlantic populations, and via purifying selection in eastern Pacific populations. Results of distribution of variation within and among populations, fixation indices and clustering analyses are very similar between regions: all regions show almost all variation within populations, which is consistent with balancing selection.

fuhctm in the northwestern Atlantic populations is undergoing weaker balancing or directional selection than at *fuhc^{sec}*, based on the ω statistic. ω identifies fewer codons with > 95% posterior probability of selection at *fuhctm* than *fuhc^{sec}*, and no exons have elevated ω values. AMOVA, fixation indices, pairwise population F_{ST} values and clustering analyses show that variation within populations at *fuhctm* is elevated with respect to neutrally evolving housekeeping genes (balancing selection). On the other hand, neutrality statistics provide evidence for purifying selection. Fu and Li's D/F and Tajima's D statistics have much more negative values when nonsynonymous sites are used than when synonymous sites are used.

Selection in *fuhctm* in the eastern Pacific populations is also weaker than at *fuhc^{sec}*. omegaMap 0.5 found half as many codons

with > 95% posterior probability of selection in *fuhctm* as *fuhc^{sec}*, despite the similar lengths of the two transcripts. The Seattle, WA population had higher ω values than the rest of the gene for a single exon or exonic region in each gene: Exon 5 in *fuhc^{sec}* and Exons 4-5 together in *fuhctm*. There is more variation within populations in *fuhctm* than in housekeeping genes. This difference is much less pronounced for *fuhctm* than for *fuhc^{sec}*, implying that balancing selection is weaker at *fuhctm* than at *fuhc^{sec}*. All three neutrality statistics point towards *fuhctm* undergoing purifying selection.

fuhctm in the English Channel populations does not have very many codons showing evidence of balancing or directional selection ($\omega > 1$), and only a single population for a single exon has elevated ω values (Exon 10 for Hamble). Like *fuhc^{sec}*, AMOVA, tests of differentiation based on allele frequencies and model-based clustering analyses provide strong evidence for balancing selection. In contrast to *fuhc^{sec}*, English populations show evidence of balancing (Fu and Li's D/F) or purifying (Tajima's D) selection while French populations are undergoing purifying selection.

Comparing all three geographic regions, selection on *fuhctm* may be stronger in the northwestern Atlantic populations than in the other regions, but selection in all three regions appears to be less pronounced in *fuhctm* than in *fuhc^{sec}*. The northwestern Atlantic populations have more codons with a high posterior probability of selection, but all three regions have many fewer codons under selection in *fuhctm* than in *fuhc^{sec}*, despite the similar lengths of the two proteins. All three regions have similar W values for Exons 3 and 10, and Exons 1-5, although the English Channel has higher values for Exons 1-3 and Exons 4-5 than the other two regions. A single population in the northwestern Atlantic has higher ω values in Exon 3

than in the rest of the transcript, a single population in the English Channel has higher ω values in Exon 10 than in the rest of the transcript, and a single population in the eastern Pacific region has higher ω values in Exon 4-5 than in the rest of the transcript.

The eastern Pacific region has more population structure at *fuhctm* than the other regions, but all three regions show evidence of balancing selection with AMOVA/fixation indices. Using Fu and Li's D and F and Tajima's D, the evolutionary history of *fuhctm* is similar across all three regions: *fuhctm* is evolving under balancing or purifying selection in English populations, and under purifying selection in French, northwestern Atlantic and eastern Pacific populations. Across geographic regions, selection is weaker at *fuhctm* than at *fuhc^{sec}*.

Expanding the focus beyond Clade A, *fuhc^{sec}* and *fuhctm* have polymorphisms that clearly transcend species boundaries. There are no fixed differences between Clade A and Clade E at either gene, when the uncorrected p distance at mtCOI is 13.8%. The same result was seen between humans and chimpanzees at four MHC genes: MHC A, MHC B, MHC C, and *DRB1* (Garrigan and Hedrick 2003), and in the promoter of a MHC Class II gene in primates (Loisel et al. 2006). Absence of fixed differences between species is evidence for balancing selection maintaining variation at these loci (Garrigan and Hedrick, 2003). Balancing selection can also be inferred from the difference in amount of lineage sorting when comparing the genealogies of the *fuhc* genes to 18S rRNA. While 18S rRNA Clade A and Clade E alleles are well-sorted, there are no *fuhc^{sec}* clades of more than five alleles that contain only Clade A or Clade E alleles. This type of genealogy, where allorecognition alleles from one species are more closely related to allorecognition alleles from another species than to other conspecific

allorecognition alleles, has been seen in SI loci in plants (Ioerger et al., 1990) and MHC Class II genes in artiodactyls (Gutierrez-Espeleta et al., 2001) and primates (Loisel et al., 2006). The synonymous substitutions within Clade A for *fuhc* genes substantially exceed the synonymous substitutions between Clade A and Clade E at four neutrally evolving genes, which is also consistent with *fuhc* alleles that are older than the speciation event between Clade A and E. Assuming the number of synonymous substitutions per site per year is equal in *fuhc* and the housekeeping genes, then the *fuhc* alleles within Clade A must have diverged in the ancestor of Clade A and Clade E. This same pattern is seen in SI loci in plants (Ioerger et al., 1990).

Either interspecific gene flow or balancing selection retaining ancestral polymorphisms could explain many of these observations relating to trans-specific polymorphisms. While allele exchange in an ancient introgression event cannot be ruled out, there are two arguments against recent gene flow. First, no recent introgression was detected in the English Channel between Clade A and Clade E using microsatellite loci (Bock et al., 2012). Second, mtCOI divergence between Clade A and Clade E matches the highest value associated with a marine broadcast spawning species pair that can still introgress (Nydam and Harrison, 2011). Clade A and Clade E, as spermcast spawning species, have more potential for reproductive barriers than externally fertilizing broadcast spawning species. Therefore, introgression between Clade A and Clade E is extremely unlikely.

The inference of balancing selection acting on the *fuhc* genes is supported by trans-species polymorphism analyses, AMOVA, fixation indices, and the ω statistics. F_{ST} statistics can be depressed in highly polymorphic loci, even if there is

significant differentiation between populations (Jost 2008). However, the AMOVA framework does not rely on allele frequencies and is therefore not subject to the same potential artifacts as F_{ST} . The codons identified as under selection by the ω statistic could be evolving under balancing or directional selection. Given the consistent signal of balancing selection in the AMOVA/ F_{ST} and trans-species polymorphism analyses, balancing selection seems most likely. But the neutrality tests did not detect evidence of balancing selection in either *fuhc^{sec}* or *fuhctm*, except for French populations at *fuhc^{sec}*. At *fuhc^{sec}*, each set of populations (English, French, eastern Pacific, northwestern Atlantic) seems to be evolving under a different type of selection, according to Fu and Li's D^*/F^* and Tajima's D . *fuhctm* populations appear to be under purifying selection, not balancing selection, across all regions.

We favor two possible explanations for the discrepancy between statistical tests, both involving negative frequency dependent selection as the mechanism of balancing selection. Negative frequency dependent selection occurs when rare alleles are favored (Grosberg, 1988). This type of selection is the most likely driver of balancing selection in *B. schlosseri* (Nydam et al., 2013b), because fusion is costly (Chadwick-Furman and Weissman, 1995, Chadwick-Furman and Weissman, 2003, Rinkevich and Weissman, 1992), and therefore individuals with rare alleles should fuse less often (De Tomaso, 2006). We tested this prediction in a single population (Santa Barbara, CA, USA), but the *fuhc* genes are so polymorphic that every allele we sequenced was rare (Nydam et al., unpublished data).

The first explanation is that recent negative frequency dependent selection could produce dramatic allele frequency changes, resulting in the

loss of ancient alleles (Fijarczyk and Babik, 2015). This has been demonstrated in simulations of MHC loci (Ejsmond and Radwan, 2011). There could be short-term dynamics that are unique to each population that obscure the long-term signature of balancing selection at the species level in Clade A. Second, Fu and Li's $D^*/D/F^*/F$ and Tajima's D statistics that identify an excess of intermediate frequency mutations are interpreted as evidence of balancing selection. But negative frequency dependent selection can cause mutational patterns that are close to patterns expected from neutral evolution (Ejsmond et al., 2010), which can be seen in *fuhc^{sec}* in English populations.

In summary, the allorecognition genes in *B. schlosseri* Clade A are experiencing balancing selection, which maintains the extensive polymorphism documented in these genes. Evidence for balancing selection comes from distribution of polymorphisms within and among populations, d_N/d_S statistics, and trans-species polymorphisms. Both *fuhc* genes exhibit higher within-population variation and lower population differentiation measures (F_{ST}) than neutral loci. The *fuhc* genes contain a substantial number of codons with $> 95\%$ posterior probability of $d_N/d_S > 1$. *fuhc^{sec}* and *fuhctm* also contain polymorphisms that are shared between Clade A and Clade E. These polymorphisms are more likely to be ancestral polymorphisms than the result of introgressive hybridization between the two clades.

Acknowledgements

This work was supported by Centre College, by a grant from the Metzger Fund for Field Studies to MLN, and by National Science Foundation (NSF) Grant IOS-0842138 to AWD. We thank Sarah Bialik, Kirsten Giesbrecht, Grant Nation, Jessica Peak and

Snjezana Rendulic for help with sampling and Rebecca Cook for help with data collection. We are grateful to the staff of the following marinas and ports, who allowed us access to their docks: Bureau du Port de Plaisance (Granville, France), Bureau du Port (Perros-Guirec, France), Club Nautic de L'Escala (L'Escala, Spain), Marina de Santander (Santander, Spain), MDL Marinas (Hamble Point, UK), Moulin Blanc Marina (Brest, France), Parkstone Bay Marina (Poole, UK), Poole Quay Boat Haven (Poole, UK), Port de Plaisance (Concarneau, France), Port de Plaisance (Roscoff, France), Porto de Ferrol (Ferrol, Spain), Porto de Sada (Sada, Spain), Premier Marinas (Falmouth and Gosport, UK), Puerto de Burela (Burela, Spain), and Tor Bay Harbour Authority (Torquay, UK). We thank one anonymous reviewer for helpful comments on the manuscript.

Funding

This work was supported by National Science Foundation [IOS-0842138]. The National Science Foundation played no role in the study design, the collection/analysis/interpretation of data, the writing of the manuscript or the decision to submit the article for publication.

References

Allen. A.M., Hiscock, S.J., 2008. Evolution and phylogeny of self-incompatibility systems in angiosperms. In: Franklin-Tong VE (ed) *Self-Incompatibility in Flowering Plants – Evolution, Diversity and Mechanisms*. Springer-Verlag, Berlin, pp 73-101.

Ban, S., Harada, Y., Yokosawa, H., Sawada, H., 2005. Highly polymorphic vitelline-coat protein HaVC80 from the ascidian, *Halocynthia aurantium*: structural analysis and involvement in self/nonself recognition during fertilization. *Dev. Biol.* 286, 440-451.

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

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

Brückner, S., Mösch, H.U., 2012. Choosing the right lifestyle: adhesion and development in *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* 36, 25-58.

Cao, P., Dey, A., Vassallo, C.N., Wall, D., 2015. How myxobacteria cooperate. *J. Mol. Biol.* 427, 3709-3721.

Cervantes, M.D., Hamilton, E.P., Xiong, J., Lawson, M.J., Yuan, D., Hadjithomas, M., Miao, W., Orias, E., 2013. Selecting one of several mating types through gene segment joining and deletion in *Tetrahymena thermophila*. *PLOS Biol* 11, e1001518.

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

Chadwick-Furman N.E., Weissman, I.L., 2003. The effects of allogeneic contact on life-history traits of the colonial ascidian *Botryllus schlosseri* in Monterey Bay. Biol. Bull. 205, 133–143.

Chaine, A.S., Schtickzelle, N., Polard, T., Huet, M., Clobert, J., 2010. Kin-based recognition and social aggregation in a ciliate. Evolution 64, 1290–1300.

Cutter, A.D., Wang, G.X., Ai, H., Peng, Y., 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., Nyholm, S.V., Palmeri, K.J., Ishizuka, K.J., Ludington, W.B., Mitchel, K., Weissman, I.L., 2005. Isolation and characterization of a protochordate histocompatibility locus. Nature 438, 454–459.

Ejsmond, M.J., Babik, W., Radwan, J., 2010. MHC allele frequency distributions under parasite-driven selection: a simulation model. BMC Evol. Biol. 10, 332.

Ejsmond, M.J., Radwan, J., 2011. MHC diversity in bottlenecked populations. *Conserv. Genet.* 12, 129-137.

Espinosa, A., Paz-y-Mino-C, G., 2012. Discrimination, crypticity and incipient taxa in *Entamoeba*. *J. Eukaryot. Microbiol.* 59, 101-104.

Excoffier, L., Lischer, H.E.L., 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.

Figuerola, F., Günter, E., Klein, J., 1988. MHC polymorphism pre-dating speciation. *Nature* 335, 265–267.

Fijarczyk, A., Babik, W., 2015. Detecting balancing selection in genomes: limits and prospects. *Mol. Ecol.* 24, 3529-3545.

Flajnik, M.F., Kasahara, M., 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* 11, 47-59.

Folmer, O., Hoeh, W., Black, M., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech.* 3, 294-299.

Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. *Genetics* 133, 693-709.

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

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

Goossens, K.V., Ielasi, F.S., Nookaew, I., Stals, I., Alonso-Sarduy, L., Daenen, L., Van Mulders, S.E., Stassen, C., Van Eijsden, R.G., Siewers, V., Delvaux, F.R., 2015. Molecular mechanism of flocculation self-recognition in yeast and its role in mating and survival. *mBio* 6, e00427-e00515.

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

Grosberg, R.K., Plachetzki, D., 2010. Marine invertebrates: genetics of colony recognition, in: Breed, M.D., Moore, J. (Eds.), *Encyclopedia of Animal Behavior*. Academic Press, Oxford, pp. 381-388.

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

Gutierrez-Espeleta, G.A., Hedrick, P.W., Kalinowski, S.T., Garrigan, D., Boyce, W.M., 2001. Is the decline of desert bighorn sheep from infectious disease the result of low MHC variation? *Heredity* 86, 439-450.

Harada, Y., Takagaki, Y., Sunagawa, M., Saito, T., Yamada, L., Taniguchi, H., Shoguchi, E., Sawada, H., 2008. Mechanism of self-sterility in a hermaphroditic chordate. *Science* 320, 548-50.

Hedrick, P.W., 2007. Balancing selection. *Curr. Biol.* 17, R230-231.

Hubisz, M.J., Falush, D., Stephens, M., Pritchard, J.K., 2009. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322-32.

Hudson, R.R., Kreitman, M., Aguadé, M., 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153-159.

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

Ioerger, T.R., Clark, A.G., Kao, T.H., 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *P. Natl. Acad. Sci. USA* 87, 9732–9735.

Iwano, M., Takayama, S., 2012. Self/non-self discrimination in angiosperm self-incompatibility. *Curr. Opin. Plant. Biol.* 15, 78-83.

Jost, L., 2008. G_{ST} and its relatives do not measure differentiation. *Mol. Ecol.* 17, 4015-4026.

Kalla, S.E., Queller, D.C., Lasagni, A., Strassmann, J.E., 2011. Kin discrimination and possible cryptic species in the social amoeba *Polysphondylium violaceum*. *BMC Evol. Biol.* 27, 11–31.

Karadge, U.B., Gosto, M., Nicotra, M.L., 2015. Allorecognition proteins in an invertebrate exhibit homophilic interactions. *Curr. Biol.* 25, 2845-2850.

Kaushik, S., Katoch, B., Nanjundiah, V., 2006. Social behaviour in genetically heterogeneous groups of *Dictyostelium giganteum*. *Behav. Ecol. Sociobiol.* 59, 521–530.

Librado, P., Rozas, J., 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinform.* 25, 1451-1452.

Loisel, D.A., Rockman, M.V., Wray, G.A., Altmann, J., Alberts, S.C., 2006. Ancient polymorphism and functional variation in the primate MHC-*DQA1* 5' cis-regulatory region. P. Natl. Acad. Sci. USA 103, 16331-16336.

Lopez-Legentil, S., Turon, X., Planes, S., 2006. Genetic structure of the star sea squirt, *Botryllus schlosseri*, introduced into southern European harbors. Mol. Ecol. 15, 3957-3967.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351, 652-654.

Mehdiabadi, N.J., Jack, C.N., Farnham, T.T., Platt, T.G., Kalla, S.E., Shaulsky, G., Queller, D.C., Strassmann, J.E., 2006. Social evolution: kin preference in a social microbe. Nature 442, 881-882.

Nydam, M.L., De Tomaso, A.W., 2011. Creation and maintenance of variation in allorecognition loci: molecular analysis in various model systems. Front. Immunol. 2, 79.

Nydam, M.L., Harrison, R.G., 2011. Introgression despite substantial divergence in a broadcast spawning marine invertebrate. Evolution 65, 429-442.

Nydam, M.L., Netuschil, N., Sanders, E., Langenbacher, A., Lewis, D.D., Taketa, D.A., Marimuthu, A., Gracey, A.Y., De Tomaso, A.W., 2013a. The candidate histocompatibility locus of a basal chordate encodes two highly polymorphic proteins. PLoS One 8, e65980.

Nydam, M.L., Taylor, A.A., De Tomaso, A.W., 2013b. Evidence for selection on a chordate histocompatibility locus. Evolution 67, 487-500.

Pandey, K.K., 1960. Incompatibility in *Abutilon* 'hybridum.' Am. J. Bot. 47, 877–883.

Pritchard, J.K. 2010. Documentation for *structure* software: Version 2.3

Puill-Stephan, E., Willis, B.L., Abrego, D., Raina, J.B., van Oppen, M.J.H., 2012.

Allorecognition maturation in the broadcast-spawning coral *Acropora millepora*. Coral Reefs 31, 1019-1028.

R Core Team, 2016. R: A language and environment for statistical computing. <https://www.R-project.org> (accessed 05.14.16)

Raftos, D. 1994. Allorecognition and humoral immunity and tunicates. Ann. NY. Acad. Sci. 712, 227–244.

Ramasamy, R.K., Ramasamy, S., Bindroo, B.B., Naik, V.G., 2014. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus 3, 1.

Raudaskoski, M., 2015. Mating-type genes and hyphal fusions in filamentous basidiomycetes. *Fungal Biol. Rev.* 29, 179-93.

Rendueles, O., Zee, P.C., Dinkelacker, I., Amherd, M., Wielgoss, S., Velicer, G.J., 2015. Rapid and widespread de novo evolution of kin discrimination. *P. Natl. Acad. Sci. USA* 112, 9076-81.

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

Romeralo, M., Escalante, R., Baldauf, S.L., 2012. Evolution and diversity of dictyostelid social amoebae. *Protist* 163, 327–343.

Saito, Y., 2013. Self and Nonself Recognition in a Marine Sponge, *Halichondria japonica* (Demospongiae). *Zool. Sci.* 30, 651-657.

Sato, K., Nishio, T., Kimura, R., Kusaba, M., Suzuki, T., Hatakeyama, K., Ockendon, D.J., Satta, Y., 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., Vekemans, X., Charlesworth, D., 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genet. Res.* 76, 51–62.

Scofield, V.L., Nagashima, L.S., 1983. Morphology and genetics of rejection reactions between oozoids from the tunicate *Botryllus schlosseri*. Biol. Bull. 165, 733–44.

Smith, K.F., Stefaniak, L., Saito, Y., Gemmill, C.E., Cary, S.C., Fidler, A.E., 2012. Increased inter-colony fusion rates are associated with reduced COI haplotype diversity in an invasive colonial ascidian *Didemnum vexillum*. PLOS ONE 7, e30473.

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

Stefanic, P., Kraigher, B., Lyons, N.A., Kolter, R., Mandic-Mulec, I., 2015. Kin discrimination between sympatric *Bacillus subtilis* isolates. P. Natl. Acad. Sci. USA 112, 14042–14047.

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

Taketa, D.A., De Tomaso, A.W., 2015. *Botryllus schlosseri* Allorecognition: Tackling the Enigma. Dev. Comp. Immunol. 48, 254–265.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* 30, 2725-2729.

Teske, P.R., Rius, M., McQuaid, C.D., Styan, C.A., Piggott, M.P., Benhissoune, S., Fuentes-Grünwald, C., Walls, K., Page, M., Attard, C.R., Cooke, G.M., 2011. "Nested" cryptic diversity in a widespread marine ecosystem engineer: a challenge for detecting biological invasions. *BMC Evol. Biol.* 11, 1.

Wenren, L.M., Sullivan, N.L., Cardarelli, L., Septer, A.N., Gibbs, K.A., 2013. Two independent pathways for self-recognition in *Proteus mirabilis* are linked by type VI-dependent export. *mBio* 4, e00374-e00413.

Westerman, E.L., Dijkstra, J.A., Harris, L.G., 2009. High natural fusion rates in a botryllid ascidian. *Mar. Biol.* 156, 2613-2619.

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

Wu, C.C., Diggle, P.K., Friedman, W.E., 2013. Kin recognition within a seed and the effect of genetic relatedness of an endosperm to its compatriot embryo on maize seed development. *P. Natl. Acad. Sci. USA* 110, 2217-22.

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

Yund, P.O., Collins, C., Johnson, S.L., 2015. Evidence of a native northwest Atlantic *COI* haplotype clade in the cryptogenic colonial ascidian *Botryllus schlosseri*. Biol. Bull. 228, 201-216.

Figure Legends

Fig. 1 STRUCTURE plots of Q, the proportion of each individual's ancestry attributed to population K. The colors in the plots represent the K populations. In plots where K = 3, there are three colors (white, black and grey). In plots where K = 2, there are two colors (white and black). In plots where K = 1, there is one color (black). Each individual in each plot is represented by a single vertical bar. Individuals with ancestry from more than one population have bars composed of more than one color. 1 = Brest, France, 2 = Concarneau, France, 3 = Hamble, England, 4 = Poole, England, 5 = Falmouth, MA, USA, 6 = Quissett, MA, USA, 7 = Sandwich, MA, USA, 8 = Monterey, CA, USA, 9 = Santa Barbara, CA, USA, 10 = Seattle, WA, USA. a) 12 nuclear housekeeping genes plus mtCOI, K = 3 b) *fuhc* genes, K = 1, c) *fuhc* genes, K = 2.

Fig. 2 Maximum Likelihood phylogeny of European *fuhctm* alleles. Alleles are labeled with the species identity determined by mitochondrial cytochrome oxidase I (mtCOI) and/or 18S rRNA. The location of each allele in the phylogeny is determined by *fuhctm*,

but the label of each allele is determined by mtCOI and/or 18S rRNA. Only nodes with > 50% bootstrap support are labeled with bootstrap values. Branch lengths are proportional to the number of substitutions per site. Population abbreviations are as follows: BR (Brest, France), BU (Burela, Spain), CU (Concarneau, France), ES (Escala, Spain), FE (Ferrol, Spain), FM (Falmouth, England), GR (Granville, France), GSM (Gosport, England), HPM (Hamble, England), PBM (Parkstone Bay, England), PG (Perros-Guirec, France), PO (Poole, England), RF (Roscoff, France), SA (Santander, Spain), SD (Sada, Spain), TQ (Torquay, England).

Fig. 3 Maximum Likelihood phylogeny of European *fuhc^{sec}* alleles. Alleles are labeled with the species identity determined by mitochondrial cytochrome oxidase I (mtCOI) and/or 18S rRNA. The location of each allele in the phylogeny is determined by *fuhc^{sec}*, but the label of each allele is determined by mtCOI and/or 18S rRNA. Only nodes with > 50% bootstrap support are labeled with bootstrap values. Branch lengths are proportional to the number of substitutions per site. Population abbreviations are as follows: BR (Brest, France), BU (Burela, Spain), CU (Concarneau, France), ES (Escala, Spain), FE (Ferrol, Spain), FM (Falmouth, England), HPM (Hamble, England), PBM (Parkstone Bay, England), PO (Poole, England), SD (Sada, Spain).

Fig. 4 Maximum Likelihood phylogeny of 18S rRNA alleles. Only nodes with > 50% bootstrap support are labeled with bootstrap values. Branch lengths are proportional to the number of substitutions per site. Population abbreviations are as follows: BR (Brest, France), CSM (Camaret-sur-mer, France), CU (Concarneau, France), FM (Falmouth, England), GR (Granville, France), GSM (Gosport, England), HPM (Hamble, England),

PG (Perros-Guirec, France), PO (Poole, England), QAB (Plymouth, England), TQ (Torquay, England).

Fig. 5 Maximum Likelihood phylogeny of mtCOI alleles. Only nodes with > 50% bootstrap support are labeled with bootstrap values. Branch lengths are proportional to the number of substitutions per site. Population abbreviations are as follows: AR (Ares, Spain), BR (Brest, France), BU (Burela, Spain), CSM (Camaret-sur-mer, France), CU (Concarneau, France), FLM (Falmouth, MA, USA), FM (Falmouth, England), GI (Gijon, Spain), GR (Granville, France), GSM (Gosport, England), HPM (Hamble, England), LA (Llastres, Spain), MR (Monterey, CA, USA), MU (Mutriku, Spain), PBM (Parkstone Bay, England), PG (Perros-Guirec, France), PO (Poole, England), QAB (Plymouth, England), QST (Quissett, MA, USA), RF (Roscoff, France), SA (Santander, Spain), SE (Seattle, WA, USA), TQ (Torquay, England).

Tables

Table 1: Population genetic tests used in this study to detect directional and balancing selection at *fuhc^{sec}* and *fuhctm*.

Test	Inference
Neutrality Statistics	Negative values = directional selection Positive values = balancing selection Values = 0 = neutral evolution (no selection)
McDonald Kreitman	Excess of nonsynonymous fixed differences = directional selection Excess of nonsynonymous polymorphisms = balancing selection Ratio of nonsynonymous to synonymous polymorphism is equal to the ratio of nonsynonymous to synonymous divergence = neutral evolution (no selection)

HKA	Difference in amount of polymorphism and divergence between two loci cannot be explained by difference in mutation rate = balancing selection Difference in amount of polymorphism and divergence between two loci can be explained by difference in mutation rate = neutral evolution (no selection)
ω (dN/dS)	Values > 1 = directional or balancing selection Values = 1 = neutral evolution (no selection)
Distribution of polymorphism	Lower polymorphism within populations and higher differentiation between populations than neutral loci = directional selection Higher polymorphism within populations and lower differentiation between populations than neutral loci = balancing selection Similar levels of polymorphism within populations and similar levels of differentiation between populations to neutral loci = neutral evolution (no selection)

Table 2: *fuhc^{sec}* and *fuhctm* alleles used in this study. NW Atlantic = northwestern Atlantic Ocean, CA = California USA, MA = Massachusetts USA, WA = Washington USA, FR = France, SP = Spain, UK = United Kingdom

fuhc^{sec}

Region	Population	# Clade A alleles	# Clade E alleles
Bay of Biscay	Burela, SP	0	2
	Ferrol, SP	0	2
	Sada, SP	0	3
English Channel, UK	Falmouth, UK	0	8
	Hamble, UK	26	0
	Parkstone, UK	0	2
	Poole, UK	25	4
English Channel, FR	Brest, FR	7	0
	Concarneau, FR	23	2
Mediterranean	Escala, SP	0	1
NW Atlantic	Falmouth, MA	17	0

	Quissett, MA	22	0
	Sandwich, MA	22	0
Pacific	Monterey, CA	8	0
	Santa Barbara, CA	16	0
	Seattle, WA	26	0

fuhctm

Region	Population	# Clade A alleles	# Clade E alleles
Bay of Biscay	Sada, SP	0	4
	Santander, SP	0	1
English Channel, UK	Falmouth, UK	0	12
	Gosport, UK	0	7
	Hamble, UK	27	0
	Parkstone, UK	0	2
	Poole, UK	19	4
	Torquay, UK	0	6
English Channel, FR	Brest, FR	10	2
	Concarneau, FR	20	1
	Granville, FR	0	3
	Perros-Guirec, FR	0	3
	Roscoff, FR	0	2
Mediterranean	Escala, SP	0	1
NW Atlantic	Falmouth, MA	17	0
	Quissett, MA	21	0
	Sandwich, MA	26	0
Pacific	Monterey, CA	10	0

Santa Barbara, CA	8	0
Seattle, WA	20	0

Table 3: Fu and Li's D*/D, F*/F and Tajima's D values for $fuhc^{sec}$ and $fuhc^{tm}$. In each column, the first value was obtained using all sites, the second value using non-synonymous sites, the third value using synonymous sites. *p value < 0.05, **p value < 0.01

$fuhc^{sec}$

Region	Fu and Li's D*	Fu and Li's F*	Tajima's D
England	-0.18/-0.09/-0.41	-0.16/-0.06/-0.38	-0.05/0.01/-0.18
France	0.76/1.04/0.17	0.78/1.06/0.20	0.46/0.59/0.18
NW Atlantic	-2.08**/-1.69*/-1.99*	-1.92**/-1.91*/-1.66*	-1.49**/-1.29*/-1.16*
Pacific	-0.88/-1.43**/-0.09	-0.93/-1.31/-0.34	-0.60/-0.39/-1.30*

$fuhc^{tm}$

Region	Fu and Li's D	Fu and Li's F	Tajima's D
England	1.89/1.63/1.95	1.00/0.58/1.19	-0.56/-1.32/-0.48
France	-0.05/-1.11/0.26	0.09/-1.13/0.47	0.55/-0.6/0.6
NW Atlantic	-1.90*/-2.55**/-0.32	-1.96*/-2.55**/-0.63	-1.21*/-1.51*/-0.70
Pacific	-1.26*/-0.27/0.81	-1.27*/-0.01/0.57	-0.72/-1.14*/-0.22

Table 4: $fuhc^{sec}$ and $fuhc^{tm}$ codons with > 95% posterior probability of $\omega > 1$, number of such codons in each population, W values of Mann-Whitney U Tests. *W values with a p value < 0.05, **W values with a p value < 0.01, ***W values with a p value < 0.001

$fuhc^{sec}$

Population	Region	# codons	Exon 5	Exon 6	Exon 8/9
------------	--------	----------	--------	--------	----------

Brest, FR	English Channel	16	12,663	7,444	11,580
Concarneau, FR	English Channel	18	12,149	7,490	12,671
Hamble, UK	English Channel	17	13,700**	9,093	11,803
Poole, UK	English Channel	17	12,740	9,077	11,689
Average	English Channel	17	12,813	8,276	11,936
Falmouth, MA	NW Atlantic	21	14,408**	9,917*	11,290
Quissett, MA	NW Atlantic	30	14,804***	8,724	12,833
Sandwich, MA	NW Atlantic	25	16,392***	10,123**	9,400
Average	NW Atlantic	25	15,201	9,588	11,174
Monterey, CA	Pacific	22	10,499	8,025	10,676
Santa Barbara, CA	Pacific	27	12,997	9,093	10,221
Seattle, WA	Pacific	35	13,574*	8,210	12,641
Average	Pacific	28	12,357	8,443	11,179

fuhctm

Population	Region	# codons	Exon 3	Exon 10	Exons 1-3	Exons 1-5	Exons 4-5
------------	--------	-------------	--------	---------	-----------	-----------	-----------

Brest, FR	English Channel	10	12,096	10,775	18,745	22,322	15,293
Concarneau, FR	English Channel	18	9,930	11,052	17,788	21,212	15,140
Hamble, UK	English Channel	9	11,927	12,345*	17,339	22,399	16,776
Poole, UK	English Channel	12	10,990	9,355	18,606	22,244	15,354
Average	English Channel	12	11,236	10,882	18,120	22,044	15,641
Falmouth, MA	NW Atlantic	20	10,379	10,899	18,678	22,222	9,825
Quissett, MA	NW Atlantic	10	14,058***	10,818	11,921	22,313	10,392
Sandwich, MA	NW Atlantic	17	10,540	10,145	18,542	22,544	10,111
Average	NW Atlantic	16	11,659	10,621	16,380	22,360	10,109
Monterey, CA	Pacific	12	11,889	10,399	17,547	23,614	11,406
Santa Barbara, CA	Pacific	18	11,312	11,453	18,416	23,704	11,035
Seattle, WA	Pacific	10	11,577	9,042	14,672	24,310	13,366*

Average	Pacific	13	11,593	10,298	16,878	23,876	11,936
---------	---------	----	--------	--------	--------	--------	--------

Table 5: *fuhc^{sec}*, *fuhctm* and housekeeping gene percentage variation within populations, F_{ST} and population pairwise F_{ST} values for Clade A. Pairwise F_{ST} values are averages of all pairwise F_{ST} values between all populations in the region. Housekeeping gene values are averages of all four housekeeping genes amplified for each region.

English Channel	<i>fuhc^{sec}</i>	<i>fuhctm</i>	Housekeeping genes
Variation within populations	98.75	98.22	80.45
F_{ST}	0.01	0.02	0.20
Population pairwise F_{ST}	0.02	0.01	0.15

NW Atlantic	<i>fuhc^{sec}</i>	<i>fuhctm</i>	Housekeeping genes
Variation within populations	94.83	100.40	86.34
F_{ST}	0.05	0.00	0.14
Population pairwise F_{ST}	0.06	0.00	0.15

Pacific	<i>fuhc^{sec}</i>	<i>fuhctm</i>	Housekeeping genes
Variation within populations	100.59	79.42	69.31
F_{ST}	-0.01	0.21	0.34
Population pairwise F_{ST}	-0.01	0.18	0.28

Table 6: A summary of inferences of selection in each region from all analyses. B = Balancing, D = Directional, P = Purifying. Evidence of strong selection is indicated by bold font.

fuhc^{sec}

Region	ω	Distribution of polymorphism	Fu and Li's D*F*	Tajima's D
England	B/D	B	No selection	No selection
France	B/D	B	B	B
NW Atlantic	B/D	B	D	D
Pacific	B/D	B	D	P

fuhctm

Region	ω	Distribution of polymorphism	Fu and Li's D*F*	Tajima's D
England	B/D	B	B	P
France	B/D	B	P	P
NW Atlantic	B/D	B	P	P
Pacific	B/D	B	P	P

Fig. 1

[illegible]

0.01

Fig. 3

Fig. 4

Fig. 5