Integrative and Comparative Biology
Integrative and Comparative Biology, volume 52, number 3, pp. 418–429
doi:10.1093/icb/ics099
Society for Integrative and Comparative Biology

SYMPOSIUM

Genetic Variation in the Acorn Barnacle from Allozymes to Population Genomics

Patrick A. Flight[1] and David M. Rand[2]

Department of Ecology and Evolutionary Biology, 80 Waterman Street, Box G-W, Brown University, Providence, RI 02912, USA

From the symposium "Barnacle Biology: Essential Aspects and Contemporary Approaches" presented at the annual meeting of the Society for Integrative and Comparative Biology, January 3–7, 2012 at Charleston, South Carolina.

[1]E-mail: Patrick_Flight@brown.edu
[2]E-mail: David_Rand@brown.edu

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

Synopsis Understanding the patterns of genetic variation within and among populations is a central problem in population and evolutionary genetics. We examine this question in the acorn barnacle, *Semibalanus balanoides*, in which the allozyme loci *Mpi* and *Gpi* have been implicated in balancing selection due to varying selective pressures at different spatial scales. We review the patterns of genetic variation at the *Mpi* locus, compare this to levels of population differentiation at mtDNA and microsatellites, and place these data in the context of genome-wide variation from high-throughput sequencing of population samples spanning the North Atlantic. Despite considerable geographic variation in the patterns of selection at the *Mpi* allozyme, this locus shows rather low levels of population differentiation at ecological and trans-oceanic scales ($F_{ST} \sim 5\%$). Pooled population sequencing was performed on samples from Rhode Island (RI), Maine (ME), and Southwold, England (UK). Analysis of more than 650 million reads identified approximately 335,000 high-quality SNPs in 19 million base pairs of the *S. balanoides* genome. Much variation is shared across the Atlantic, but there are significant examples of strong population differentiation among samples from RI, ME, and UK. An F_{ST} outlier screen of more than 22,000 contigs provided a genome-wide context for interpretation of earlier studies on allozymes, mtDNA, and microsatellites. F_{ST} values for allozymes, mtDNA and microsatellites are close to the genome-wide average for random SNPs, with the exception of the trans-Atlantic F_{ST} for mtDNA. The majority of F_{ST} outliers were unique between individual pairs of populations, but some genes show shared patterns of excess differentiation. These data indicate that gene flow is high, that selection is strong on a subset of genes, and that a variety of genes are experiencing diversifying selection at large spatial scales. This survey of polymorphism in *S. balanoides* provides a number of genomic tools that promise to make this a powerful model for ecological genomics of the rocky intertidal.

Introduction

The patterns of genetic variation in natural populations are fundamental to the study of evolution. The levels of standing polymorphism within populations and the degree of differentiation among geographic locations provide basic information about the balance of forces shaping the genetic basis of evolutionary change. The acorn barnacle *Semibalanus balanoides* is a wonderful model to dissect this problem as it offers a life history well-suited for the analysis of evolutionary forces in the wild. With a circumboreal distribution, wide dispersal of pelagic larvae, large population densities following recruitment to the rocky intertidal, and a sessile habit that enforces a commitment to life in a given microhabitat, *S. balanoides* experiences environmental stressors at multiple spatial scales (Barnes 1953; Barnes and Barnes 1954; Southward and Crisp 1954; Wethey 1983, 1984). It follows intuition that organisms living in heterogeneous environments will harbor more genetic variation (Maynard Smith 1998), and the high heterozygosities of *Mpi* and *Gpi* in marine organisms have long been suspect in light of this intuition. Early surveys of allozymes described general patterns of variation around the Atlantic shorelines, suggesting racial differentiation

Advanced Access publication July 4, 2012
© The Author 2012. Published by Oxford University Press on behalf of the Society for Integrative and Comparative Biology. All rights reserved. For permissions please email: journals.permissions@oup.com.

due to glacial vicariance or limited gene flow across the Atlantic (Flowerdew and Crisp 1975, 1976; Flowerdew 1983). More recently, a number of studies have identified the allozyme loci Mpi and Gpi as targets of balancing selection across different spatial scales (Holm and Bourget 1994; Schmidt and Rand 1999, 2001; Veliz et al. 2004, 2006). These studies show that frequencies of *Mpi* and *Gpi* alleles can vary as much or more between tidal levels or small geographic scales than across the Atlantic, implying that selection plays some role in the presumed racial differences. Studies of microsatellites have confirmed and extended these patterns of ecological differentiation through comparisons to nonallozyme loci (Dufresne et al. 2002), and revealed patterns of isolation by distance in the North Atlantic (Flight et al. 2012). The phylogeographic history of *S. balanoides* in the North Atlantic involves postglacial expansion and episodes of trans-Atlantic colonization that imply a dynamic history over past millennia (Brown et al. 2001; Wares and Cunningham 2001; Flight et al. 2012). While these studies have provided clear examples of fundamental questions in ecological and evolutionary genetics using the standard battery of allozymes, mtDNA, and microsatellites, our understanding of genetic variation in *S. balanoides* is based on remarkably few markers (two allozymes, mtDNA, and three to five microsatellite markers).

It is now straightforward to survey large numbers of markers across a genome to quantify variation within and between populations. Advances in high-throughput sequencing technologies have provided new tools for asking questions about gene flow and local adaptation. Typically, F_{ST} values, or a similar metric of differentiation, are calculated from a random sample of the genome and the distribution of F_{ST} values is used as a genome-wide control for the demographic history of the populations (Lewontin and Krakauer 1973; Beaumont and Nichols 1996; Beaumont and Balding 2004; Beaumont 2005). Outliers from the distribution are presumed to be either directly under selection, or closely linked to a selected locus. The combined application of high-throughput sequencing and F_{ST} outlier analysis allows unbiased screens of the genome without prior knowledge of which candidate genes to choose. This ''reverse ecology'' approach is a promising means of identifying genes that are the target of natural selection, especially when the samples being compared are stratified by well-understood geographic or ecological variables (Wood et al. 2008; Hohenlohe et al. 2010). Here we place the patterns of variation at the allozyme Mpi in the context of genome-wide variation at the nucleotide level with a population genomic survey of *S. balanoides.*

The *mannose-6-phosphate isomerase* (*Mpi*) and the *glucose phosphate isomerase (Gpi)* genes in *S. balanoides* both show a common allozyme polymorphism in which the slow and fast electromorphs are selectively linked to environmental variables on multiple spatial scales (Holm and Bourget 1994; Schmidt and Rand 1999; Schmidt et al. 2000; Schmidt 2001; Schmidt and Rand 2001; Brind'Amour et al. 2002; Veliz et al. 2004, 2006; Flight et al. 2010). However, the patterns of variation are quite different in different geographic locations. In the estuaries of the ME coast, selection is associated with tidal height and thermal/desiccation stress, with the MPI-fast allele being favored in stressful microhabitats and the MPI-slow allele favored in benign microhabitats (Schmidt and Rand 1999, 2001). The GPI allozyme in ME is robustly neutral with respect to microhabitat associations. In the Gulf of St. Lawrence, the fine-scale tidal-height selection is less apparent and the selection gradient is evident at a meso-scale spanning the mouth of the Miramichi estuary (Veliz et al. 2004, 2006). Moreover, in the Miramichi region, habitat-specific selection is operating on both the MPI and GPI allozymes. In Narragansett Bay, RI, selection is associated both with tidal-height microhabitats and with meso-scale upper-bay versus open-coast habitats. The selection is distinct from that in ME as the *Mpi* zonation is reversed, and *Gpi* does vary with tidal height (Rand et al. 2002). The evidence for genotypic trade-offs or crossing allelic fitness values at *Mpi* and has led multiple authors to suggest that balancing selection is maintaining genetic variation at the locus (Schmidt et al. 2000; Veliz et al. 2006; Flight et al. 2010).

Inferences about the patterns of selection at Mpi, and the variation in selection response in different geographic populations, have been made by contrasts to Gpi, mtDNA, or allozymes. In short, the same logic of the F_{ST} outlier approach has been applied, but very few loci have been used for contrast between loci (Schmidt and Rand 1999; Dufresne et al. 2002). The different patterns of selection in different geographic locations has begged the question of limited gene flow and population differentiation. Similarly, this question has been addressed with the same small sample of loci (fewer than eight). Here we address these questions at a genome-wide scale by performing genomic screens of three populations using high-throughput sequencing. Libraries were created by pooling genomic DNA from 20 individuals each per population from RI, ME, and UK and sequenced with the Illumina

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

GAIIx and HiSeq chemistries (San Diego, California, USA). We used these data to create a draft genome of *S. balanoides*, the first for a cirripede, identify high-quality single-nucleotide polymorphisms (SNPs), and perform F_{ST} screens. Additionally, unlike traditional genomic screening methods that generally rely on anonymous markers (e.g., AFLPs or microsatellites), sequence libraries from nonmodel organisms allow markers to be curated by homology to genes in existing genome projects through BLAST annotation. F_{ST} outlier analyses of genomic regions containing defined BLAST homology increases the opportunity to interpret the functional significance of markers that show elevated population differentiation consistent with a history of selection. The results provide the most complete picture to date of genetic variation in *Semibalanus*, and identify gene regions for future studies of natural selection at multiple spatial scales.

Materials and methods

Allele frequency heterogeneity of the *Mpi* polymorphism

Data on *Mpi* and *Gpi* allele frequencies were collected from published reports listed in Table 1. F_{ST} values were calculated as a standard two-allele conditions based on reported allele (p, q) or genotype frequencies for the MPI-fast/MPI-slow and the GPI-fast/GPI-slow allozymes: $F_{ST} = (H_{TOTAL} - H_{SUB})/H_{TOTAL}$, where H_{TOTAL} is the heterozygosity, $H = 2pq$ for the mean frequency of each allele across samples, and H_{SUB} is the mean value of $H = 2pq$ among the subpopulations that comprise the total sample. The nucleotide basis of the Mpi fast/slow polymorphism has been identified as a charge altering amino-acid polymorphism near the carboxy terminal of the protein (Flight 2011). The details will be reported elsewhere, but the molecular data confirm that the allozyme is a reliable Mendelian marker and shows 95% correspondence with the SNP causing the change in amino acids. Thus, patterns of allelic variation and F_{ST} for the Mpi allozyme provide an accurate measure of nucleotide variation at this SNP, and can be compared directly to other nuclear SNPs described below.

Illumina library sequencing

DNA from 20 barnacles from each of the three sites (RI, ME, and UK) was extracted with a Qiagen DNeasy tissue kit according to manufacturer's instructions. DNA was run on a 2% gel to inspect quality and quantified using a Quant-it broad range fluorescence kit (Invitrogen) in a SpectraMax M5 microplate reader (Molecular Devices, Sunnyvale, CA). An equal amount of DNA from each of the 20 individuals per site was combined into a pooled sample, resulting in a single pooled library per site. This approach has been shown to be effective when the number of alleles in the pool is greater than the average sequence coverage (Futschik and Schlotterer 2010; Kolaczkowski et al. 2011). The pooled sample was treated with RNAse A (Qiagen, Valencia, CA) to remove RNA contamination. DNA was sheared using DNA fragmentase enzyme from New England Biolabs (New England BioLabs, Ipswich, MA). Samples were digested for 20 min at 37°C and a band corresponding to 350–400 bp was excised from a 2% low-range agarose gel (Bio-Rad, Hercules, CA). Preparation of the samples continued following the NEBnext protocol according to manufacturers instructions. The final libraries were gel excised and run on an Agilent Bioanalyzer (Santa Clara, CA) to assess quality. The library from UK was sequenced using 100-bp paired-end (PE) reads in two lanes of an Illumina GAIIx sequencer. Each of the three libraries (RI, ME, UK), was also sequenced in an individual lane of the Illumina HiSeq sequencer as 100-bp PE reads.

Genomic assembly and screens

The resulting libraries had the last 15 bp of each 100-bp read removed due to diminished quality. They were further screened for quality using the following criteria: (1) sequences that did not pass the Illumina filter were excluded; (2) sequences were excluded if they had more than five low-quality bases as determined by an Illumina ascii score of "B"; (3) sequences were excluded if the mean quality score across the whole sequence was less than 30—corresponding to an error rate of 0.001 (scripts for quality control were modified from versions kindly provided by A. Reich, Brown University). The resulting sequences were assembled with SOAPdenovo (Li et al. 2010) using a kmer length of 31. The "M" flag in the assembly was also set to 3 due to presence of multiple individuals in the sequencing pools. This results in a draft genome assembly of many individual contigs.

For realignment and annotation all contigs longer than 1 kb from the SOAPdenovo assembly were used as the "reference genome." Each of the contigs was blasted against metazoans in a local download of the "NR" database using the BLASTX algorithm (http://blast.ncbi.nlm.nih.gov/Blast.cgi; BLAST scripts were modified from versions kindly provided

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

Table 1 Mpi and Gpi allele frequencies

Mpi and Gpi allele frequencies and population structure				
Flowerdew (1983)	MPI-slow	F_{ST}	GPI-fast	F_{ST}
North America				
Indian river	0.344		0.277	
Coney island	0.319		0.271	
Boston	0.412		0.308	
Mount desert island	0.393	0.0060	0.270	0.0012
Newfoundland	0.207		0.245	
Iceland, Hvalfjordur Fjord	0.187		0.254	
Iceland, Fragriskogur	0.188	0.0005	0.215	0.0015
Shetland	0.451		0.412	
Norway	0.452		0.427	
Denmark, Kyndby-vaerket	0.491		0.357	
Denmark, Lumsär & Overby	0.557		0.438	
Denmark, Abrena	0.500		0.415	
Holland	0.454	0.0058	0.361	0.0041
British Isles				
Carnoustie	0.521		0.458	
Loch Carron	0.553		0.395	
Culllercoats	0.516		0.359	
Robin Hood's Bay	0.531		0.407	
Menai Straits	0.505		0.432	
Bundoran	0.518		0.394	
Balbriggan	0.548		0.387	
Kilkee	0.510		0.388	
Hastings	0.465	0.0024	0.495	0.0064
France				
Cap Gris Nez	0.532		0.575	
Brest	0.491	0.0017	0.402	0.0299
North Atlantic	0.444	0.0487	0.373	0.0292
Holm and Bourget (1994)				
Nuuk E	0.228		0.495	
Nuuk P	0.257		0.513	
Iqaluit E	0.273		0.392	
Iqaluit P	0.191		0.466	
Saint Augustin	0.206		0.268	
Capucins	0.272		0.189	
Port Elgin	0.539		0.478	
Ingonish	0.472		0.385	
Mort Morien	0.424		0.368	
Queensland Beach	0.395		0.347	
Saint Andrews	0.412	0.0575	0.203	0.0502
Port Daniel	0.270		0.293	
Shippegan	0.250		0.239	
Burnt Church	0.272		0.309	
Cap Lumiere	0.484		0.447	
Shediac	0.495		0.436	
Port Elgin	0.495		0.427	
Pictou	0.444		0.392	
Port Hood	0.456		0.473	
Cheticamp	0.469	0.0418	0.443	0.0548
Miramichi North	0.264		0.280	
Miramichi South	0.474	0.0473	0.436	0.0265
Schmidt and Rand (1999, 2001)				
Hot site high 1994	0.305		0.275	
Hot site low 1994	0.415	0.0131	0.232	0.0024
High exposed cohort	0.323		0.292	
Low Algae cohort	0.395	0.0056	0.260	0.0013
Schmidt et al. (2000)				
High Exposed transplant	0.254		0.260	
Low Algae transplant	0.417	0.0298	0.272	0.0002
Veliz et al. (2004)				
Burnt Church	0.320		0.304	
Cap Lumiere	0.480	0.0267	0.438	0.0192
Veliz et al. (2006)				
Miramichi North 2001	0.318		0.262	
Miramichi South 2001	0.476	0.0260	0.415	0.0259
Miramichi North 2000	0.329		0.305	
Miramichi South 2000	0.486	0.0255	0.497	0.0382

The values reported are for the MPI-slow and the GPI-fast allozyme alleles, and Fst values are reported assuming a simple two-allele polymorphism for each locus. The F_{ST} value for each set of population samples is displayed in the column adjacent to the respective allele. Data were tabulated from the references cited in the table.

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

by M. Howison, Brown University). A conservative threshold of 10^{-10} was set for a contig to be considered a coding region. The best-scoring open-reading frame was selected from among the BLAST hits to annotate base positions in the barnacle contig.

SNP identification

To identify SNPs, the complete set of sequence reads that were used to build the genome assembly were realigned back to the reference genome of contigs using Bowtie (Langmead et al. 2009) with a seed length of 24 and up to three mismatches allowed in the seed. Sequences were aligned as single ends due to the structure of the contigs, which may have included one end of a PE read, but not the other end. Alignments were only considered if they had a single best match to the reference contigs. Other settings were the default in Bowtie. A pileup file was created using SAMtools (Li et al. 2009) and sequence variants (SNPs) were called with a custom Python script, as follows. Any position with a depth of coverage outside a predetermined range (6–35×) was excluded from the analysis to reduce the impact of low coverage and to avoid inclusion of paralogus loci or repetitive regions in the SNPs attributed to a single locus. Singletons were not considered in subsequent analyses because they could not be distinguished from sequencing errors. Furthermore, any nucleotide site with more than two alleles and any insertions or deletions, which do not map in Bowtie, were not considered in further analyses. For those contigs with strong BLAST hits, SNPs were tabulated by codon position in the best-scoring open-reading frame. Patterns of mutation were inferred using the consensus nucleotide as the reference state and the variable nucleotide as the mutation. The end result was a sample of more than 335,000 nonsingleton SNPs, about 5% of which were known to lie in regions of the barnacle genome showing homology to protein-coding genes in GenBank.

F_{ST} outlier analyses

F_{ST} estimates for the contigs were made using the unbiased method described by Kolaczkowski et al. (2011). Estimates were made on a per SNP basis and averaged across each contig. Individual SNPs that yielded a negative F_{ST} were set to zero prior to averaging. Following the approach described by Kolaczkowski et al. (2011), empirical F_{ST} outliers in pairwise comparisons were determined by taking the 1% tail of the F_{ST} distribution. While outliers can also be defined using software that seeks to estimate a null distribution of F_{ST} values based on a model of genetic drift, e.g., DetSel (Vitalis et al. 2003) or Fdist, Fdist2 (Beaumont and Nichols 1996), the 1%-outlier approach is potentially more objective. The model-based approaches are sensitive to the models of genetic drift and population structure which is further complicated when singletons are ignored in the F_{ST} estimation. Moreover, *S. balanoides* has experienced population expansions on both sides of the Atlantic, so a multitude of models could be explored to generate a variety of null distributions. The 1%-outlier approach represents an objective cut off that can serve as a benchmark for validation across F_{ST} distributions from different pairs of populations (Kolaczkowski et al. 2011).

Results

Population genetics of Mpi

Table 1 shows frequencies of MPI-slow and GPI-fast alleles in different geographic locations as tabulated from the literature. In each case, population subdivision (F_{ST}) is estimated for different samples as a measure of the degree of differentiation. Despite significant and repeatable differences in allele frequencies, F_{ST} values are rather low, on the order of a few percent. For the entire North Atlantic (Flowerdew 1983) the Mpi $F_{ST} = 0.047$, while $F_{ST} = 0.057$ for Greenland to the Bay of Fundy (Holm and Bourget 1994). For pairs of well-differentiated sites, such as across the Miramichi estuary (Holm and Bourget 1994; Veliz et al. 2004, 2006), or between thermally stressed tidal microhabitats on the ME coast (Schmidt et al. 2000), F_{ST} values are less than 4%. These finer-scale differences can arise each year due to genotype-specific mortality after settlement. The inter-population F_{ST} values are presumably more stable, but depend critically on which microhabitats are sampled. Nonetheless, for these allozyme markers that are known to be under strong ecological selection, the F_{ST} values are relatively small and, as reported below, are close to the median F_{ST} values estimated from more than 22,000 markers in the population genomic screen.

Genomic assembly and remapping

Table 2 summarizes the descriptive statistics for the *S. balanoides* genome project. After quality control 6.59×10^8 reads were used to build an assembly of the *S. balanoides* genome in SOAPdenovo. Each read was 85-bp long resulting in just over 56 Gb in total. The size of the haploid genome of *S. cariosus*, the sister species to *S. balanoides*, has been estimated at 1.37 Gb (Bachmann and Rheinsmith 1973; Gregory 2011). Assuming no dramatic change in genome

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

Table 2 Descriptive statistics for the genome assembly of *S. balanoides*

Genome size	1,370,000,000
Sequence reads (85 bp)	659,000,000
Total sequence	56,015,000,000
Average coverage per base	40.89
N50 for contigs	250
Contigs > 1 kb	22,986
Mean (median) contig length	1361 (1225)
Genome size of assembled contigs	31,291,816
Sample of assembly for SNP discovery	19,472,346
All SNPs	1,650,000
SNPs excluding singletons	335,867
Proportion of polymorphic sites (all SNPs)	0.0847
Proportion of polymorphic sites (no singletons)	0.0172

See text for details on specific terms.

Table 3 Nucleotide variation in coding and noncoding positions

Nonsingleton SNPs	Counts	Percent coding	Percent total
First codon position	3812	20.42	1.13
Second codon position	3328	17.83	0.99
Third codon position	11,527	61.75	3.43
"Noncoding"[a]	317,200		94.44
Total	335,867		

[a]Contigs not passing the BLAST threshold of $e < 10^{-10}$.

size since these species diverged, we have approximately 41× coverage of the genome of *S. balanoides* in the dataset. With 120 alleles in the total sample (3 pooled population samples × 20 individuals per sample × 2 alleles per diploid), this ratio of alleles to sequence coverage reduces the sampling effects that could occur during preparation of the library and generation of DNA sequences (Futschik and Schlotterer 2010). The N50 for the contigs in the assembly was 250 bp and 39 million contigs longer than the kmer were recovered. Of the 39 million contigs, 22,986 were 1 kb or longer and these were used as the "reference genome," with 4236 contigs displaying significant homology to a protein-coding gene at $e < 10^{-10}$. Contigs totaled 31,291,816 bases, which is approximately 2.29% of the total genome. The mean length of the reference contigs was 1361 bp (median 1225 bp).

For the Narragansett library, 2.46% of the approximately 230 million reads remapped to the genome with a unique best match. For the Harpswell and Southwold libraries 2.40% of 250 million reads and 2.29% of 180 million reads mapped, respectively. Based on these values, average coverage levels were 15× for Narragansett, 16× for Harpswell, and 11× for Southwold. Of the 31.3 million base pairs of aligned sequences, 19,472,346 were screened for variation based on our predetermined criteria of a minimum of 6× and a maximum of 35× coverage in each population. Approximately 1,650,000 variable sites were found; however, since singletons could not be distinguished from sequencing errors they were excluded from the dataset, resulting in 335,867 nonsingleton SNPs (Table 2). Ratios of these values (1,650,000/19,472,346 and 335,867/19,472,346) suggest approximate upper and lower bound estimates of the proportion of polymorphic nucleotide sites in the *S. balanoides* genome of 0.085–0.017 (Table 2). The upper value is clearly an overestimate as it includes singletons, some of which are sequencing errors; the lower value is probably an underestimate as it excludes all singletons. Despite these uncertainties, the values are similar to estimates of nucleotide heterozygosity in *Drosophila simulans* of 0.0135–0.0180 for X-linked or autosomal sequences, respectively (Begun et al. 2007).

Nucleotide variation in protein-coding genes

The quality of the assembly and remapping process are critical issues for interpreting the information content of SNPs discovered in population genomic screens. To address these questions, the 4236 contigs with strong BLAST homology to protein-coding genes were screened for SNPs in each of the three codon positions of the highest-scoring open-reading frames. As predicted from functional constraints, third codon positions showed the most variation, followed by first and second codon positions (Table 3). This pattern is well known; for example, polymorphism data at first, second and third codon positions in *Anopheles* mosquitoes show the following percentages of variation: 23.1%, 13.1%, 63.7% (Wondji et al. 2007), very similar to our data from *S. balanoides* (Table 3). In addition, the transition/transversion (ti/tv) ratio, defined as twice the number of observed transitions versus the observed transversions, was consistent with known patterns of transition bias in protein-coding genes; the third codon position showed the greatest transition bias (4.44), and the second position the least (2.19). Across all nonsingleton SNPs (both coding and noncoding), the observed transition/transversion ratio was 3.07:1, consistent with a mixture of coding and noncoding DNA (Table 4). These patterns indicate that the assembly and SNP discovery pipeline applied

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

Table 4 Mutation patterns in coding and noncoding DNA

First codon position		A	C	G	T
ti/tv	A	–	267	380	101
2.78	C	242	–	156	705
	G	784	184	–	251
	T	112	349	281	–
Second codon position		A	C	G	T
ti/tv	A	–	385	367	141
2.19	C	161	–	107	506
	G	483	132	–	160
	T	146	385	355	–
Third codon position		A	C	G	T
ti/tv	A	–	357	1087	208
4.44	C	650	–	606	2903
	G	2914	576	–	659
	T	190	1043	334	–
"Noncoding"[a]		A	C	G	T
ti/tv	A	–	16,168	40,016	15,738
3.05	C	20,466	–	10,556	55,863
	G	55,652	10,584	–	20,282
	T	15,762	39,953	16,160	–
Total		A	C	G	T
ti/tv	A	–	17,177	41,850	16,188
3.07	C	21,519	–	11,425	59,977
	G	59,833	11,476	–	21,352
	T	16,210	41,730	17,130	–

Reference base in rows, mutated base in columns. ti/tv = transition/transversion ratio; see text. [a]Contigs not passing the BLAST threshold of $e < 10^{-10}$.

to the *S. balanoides* genomic sequence generates data consistent with patterns of nucleotide variation in other eukaryotic genomes.

F_{ST} distributions

F_{ST} values were determined for each of the 22,986 contigs that were 1 kb or longer, including the 4236 contigs with homology to a protein-coding gene at $e < 10^{-10}$. The distributions for these F_{ST} values in each pair of populations are shown in Fig. 1. In some cases, there were no SNPs in the contigs for the pairwise comparisons between two sites. These cases were excluded from the distributions. Median pairwise F_{ST} between Narragansett, RI and Southwold UK was 0.0408, with a 1% tail beginning at 0.2114. Between Harpswell, ME and Southwold, UK the corresponding values were 0.0362 and 0.1718; and between Narragansett, RI and Harpswell, ME they were 0.0243 and 0.1544. The numbers of contigs in the 1% tail for each F_{ST} comparison, and those shared in different pairs of comparisons, are shown in Fig. 2. The two F_{ST} comparisons involving the UK share ∼45–58% of their outliers (RI versus UK = 102; ME versus UK = 131; with 59 shared). The two F_{ST} comparisons involving RI share ∼35–48% of their outliers (RI versus UK = 102; RI versus ME = 140; with 49 shared). The two F_{ST} comparisons involving ME share the fewest F_{ST} outliers: ∼14–15%. There were two loci that were outliers in all three distributions. One locus had significant homology to a hypothetical protein in *Tribolium* (GenBank accession: EFA10857) and may be nicotinate phosphoribosyltransferase based on homology in other species. The other locus was an unannotated transcript. Interestingly, a protein with significant homology to a settlement-inducing complex in *Balanus amphitrite* (Dreanno et al. 2006) was an outlier in the Harpswell versus Narragansett comparison (F_{ST} = 0.1869), but not in Narragansett versus Southwold (F_{ST} = 0.0931) or Harpswell versus Southwold (F_{ST} = 0.0864). F_{ST} values for nucleotide variation at Mpi is very close to the medians of the distributions (Flight 2011) and these values are similar to allozyme F_{ST} for Mpi (Fig. 1). A complete list of the outlying loci along with sequence data for each comparison is available upon request from the authors.

Discussion

Studies of genetic variation in natural population have advanced in lock step with the advent of novel technologies for distinguishing allelic variation. Allozymes (Lewontin and Hubby 1966), mtDNA (Avise et al. 1979), microsatellites (Schlotterer and Pemberton 1994; Goldstein and Clark 1995), AFLPs (Mackill et al. 1996), RAD tagging (Miller et al. 2007) among other markers, have provided many new insights into population genetics and evolution. The recent advances in high-throughput DNA sequencing have promised to transform traditionally nonmodel organisms (those without a genome or a community of researches focused on common genetic questions) into modern-day models for any number of questions spanning genetics through ecology (e.g., Baxter et al. 2011). The aim of the present study was to describe the draft genome sequence for the acorn barnacle *Semibalanus balanoides* and provide analyses that may contribute to future studies of

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

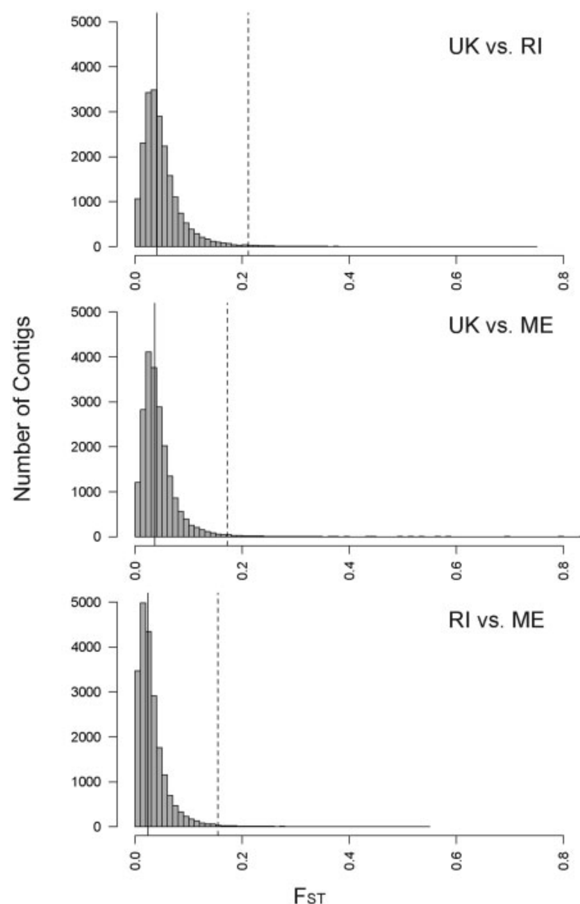

Fig. 1 Distribution of F_{ST} values for 22,986 contigs of the *S. balanoides* genome. Distributions are presented for three pairwise comparisons, as labeled on each plot. The dashed vertical line in each plot marks the 1% tail of the distribution, an objective cutoff for loci showing elevated levels of population differentiation (Kolaczkowski et al. 2011). The unbroken vertical line in each plot marks the median FST value for all contigs between each pair of populations (see Methods section). F_{ST} values for mtDNA are <0.05 for ME versus RI, 0.30 between ME and UK, and 0.29 between RI and UK (Flight et al. 2012).

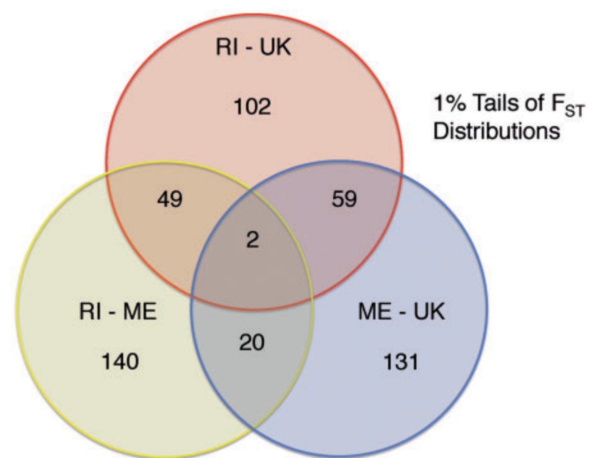

Fig. 2 Number of shared and unique contigs showing elevated F_{ST} between barnacle populations. The Venn diagram shows the number of contigs in the 1% tail of the F_{ST} distribution for each pair of populations. The overlapping sectors of each circle lists the number of shared contigs for those two pairs of F_{ST} comparisons. Thus, the RI–UK and the RI–ME comparisons share 49 contigs with elevated F_{ST}, while only two contigs are shared among all three pairwise comparisons.

genetic variation in this species. *Semibalanus balanoides* has been a common model organism among ecologists for decades, but has lagged far behind in the fields of population and evolutionary genetics. While considerable attention has been focused on the ecological genetics of the allozyme loci Mpi and Gpi in *S. balanoides* (Flowerdew 1983; Holm and Bourget 1994; Schmidt and Rand 1999; Veliz et al. 2004), and on mtDNA-based phylogeography (Wares and Cunningham 2001; Flight et al. 2012), studies of genetic variation in this species have been restricted to only a few markers. Using a pooled sequencing approach (e.g., Futschik and Schlotterer 2010; Kolaczkowski et al. 2011) and an analysis pipeline for filtering uninformative sites, we have identified more than 335,000 SNPs in thousands of anonymous contigs and open reading frames. These data allowed us to take an F_{ST} outlier approach to evaluate population substructure at more than 22,000 genomic markers. The results provide a context for the allozyme data on Mpi and Gpi, which have been interpreted as genes under balancing selection in natural populations. Together, these approaches identify a number of new studies that can be conducted to discover additional loci that show signatures of natural selection in the wild. In keeping with the goals of the Symposium, this represents both an Essential Component and a Contemporary Approach to barnacle biology.

Patterns of genetic variation in *Semibalanus*

The summary of variation at the Mpi and Gpi presented in Table 1 confirms that these loci are broadly polymorphic, show clear differences between localities at different spatial scales, and that the variation in allele frequencies can be greater at small scales than across great distances (Holm and Bourget 1994; Schmidt and Rand 1999; Veliz et al. 2004). The F_{ST} values for these allozymes generally do not exceed 5%, even spanning the North Atlantic. Studies of microsatellites further show that levels of population subdivision among North American localities are smaller than for Mpi and Gpi, and rarely exceed 2% (Dufresne et al. 2002; Flight et al. 2012).

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

The F_{ST} value for microsatellites between North America and the UK is not noticeably higher: 0.021 (Flight et al. 2012). Population subdivision for mtDNA among North American localities is 0.0445, but is substantially higher for the trans-Atlantic comparisons (~0.23–0.33) (Flight et al. 2012).

Each of these marker types has their own limitations and beg questions of the overall patterns of variation and subdivision across the barnacle genome. The allozyme data are likely modified by selection, with some combination of balancing and diversifying selection at different spatial scales. If balancing selection is a general force at Mpi and Gpi, this should prevent population differentiation leading to low F_{ST} values. The documented habitat-specific differences in allele frequency at these allozymes appears not to be strong enough to stand as an F_{ST} outlier (Table 1 and Fig. 1). The high mutation rate of microsatellites can lead to low F_{ST} values for several reasons. First, with many alleles segregating, and with a high ratio of mutation to migration rates, the interpopulation component of total variation can be a proportionally smaller component than for markers with fewer alleles or lower mutation rates (Jost 2008; Whitlock 2011). Second, high mutation rates of microsatellites can cause reversals of allelic state (homoplasy), leading to high estimates of heterozygosity and potentially masking population subdivision. MtDNA has a high mutation rate and may be subject to some of the biases discussed above for microsatellites, but the strong differentiation between Europe and North America (Wares and Cunningham 2001; Flight et al. 2012) does indicate reduced rates of gene flow relative to those among North American localities (Brown et al. 2001; Flight et al. 2012). However, conflicting evidence for different kinds of selection on mtDNA (Rand and Kann 1996; Bazin et al. 2006; Meiklejohn et al. 2007; Wares 2010), coupled with little or no recombination, suggests that mtDNA provides a limited view of the genome-wide patterns of genetic variation.

Data on single-nucleotide polymorphisms from many loci across the nuclear genome should provide a more complete picture of overall patterns of gene flow, drift, and potential locus-specific selection. For example, if the high F_{ST} value for mtDNA in trans-Atlantic comparisons (~0.30) (Flight et al. 2012) is taken as evidence for limited gene exchange across the ocean, then one would have to discount the microsatellite data from these same samples as being biased by elevated mutation rates. Moreover, if microsatellites failed to capture trans-Atlantic differentiation one could argue further that some fraction of the 1% F_{ST} outliers for nuclear SNPs in Fig. 1 may be due to reduced gene flow, consistent with the mtDNA data. Indeed, the mode and 1% cut off for F_{ST} values at >22,000 nuclear markers are higher for trans-Atlantic comparisons than for the comparison between ME and RI (Fig. 1). Such a conclusion, however, would require a double standard: that mtDNA falls among the nuclear F_{ST} outliers for the trans-Atlantic comparison, but is completely consistent with the average F_{ST} for the North American populations (see Fig. 1 and its caption). This is not a null or neutral prediction.

By considering all three comparisons of populations (Figs. 1 and 2), the data presented here do indeed provide a more complete picture of genetic variation in *S. balanoides*, and allow the allozyme, mtDNA and microsatellite data to be placed in a genome-wide context. The fact that these latter markers show limited population differentiation between ME and RI (F_{ST} values < 5%; Table 1; also see Flight et al. 2012), but dozens of nuclear markers show elevated population differentiation, strongly suggest some loci are linked to selective processes that differ between the Gulf of ME and RI. That the median of F_{ST} values in the North Atlantic is so low further points to extensive gene flow in *S. balanoides*, thereby strengthening the case for some loci under selection. Even if some fraction of these F_{ST} outlier loci are due to intragenomic effects, such as paralogous genes or repetitive DNA, not filtered by our assembly and annotation pipeline, these loci are still very interesting; population differentiation for variation in copy number among paralogs in the face of high gene flow would imply strong selection and warrant further study.

Reconciliation of Mpi and genomics

The evidence for selection on the Mpi polymorphism has come largely from repeatable differences in allele frequency across ecological gradients (Holm and Bourget 1994; Schmidt and Rand 1999), and from cohort analyses that reveal repeatable shifts in allele frequency across time (Schmidt and Rand, 2001; Brind'Amour et al. 2002; Veliz et al. 2006; Flight et al. 2010). The argument for balancing selection has come from opposing selection coefficients for alternative alleles or genotypes in alternative habitats. These real-time studies can generate stable equilibria consistent with models of balancing selection (Schmidt et al. 2000; Veliz et al. 2006) but cannot provide insight into the historical nature of the selection. Moreover, the patterns of selection on Mpi and Gpi show significant differences among geographic locations from RI to ME and the Gulf

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

of St. Lawrence (Schmidt and Rand 2001; Rand et al. 2002; Veliz et al. 2004; Flight et al. 2010). Two questions have emerged that may account for these population-specific patterns of ecological selection. First, the Mpi allozyme polymorphism may have a distinct genetic basis in the different geographic locations; second, gene flow among these locations may be sufficiently low that local adaptation has led to different genetic backgrounds on which selection at Mpi may act (Flowerdew 1983; Bertness and Gaines 1993; Holm and Bourget 1994).

The first question has been resolved by sequence analysis of the Mpi locus (Flight 2011). The Mpi fast–slow allozyme polymorphism is due to a SNP causing an amino-acid charge change showing 95% correspondence with the protein electromorphs in Rhode Island, and the same polymorphism is found in RI, ME, and the UK (Flight 2011). Evidence from detailed nucleotide sequences in support of historical balancing selection at Mpi is beyond the scope of this study and will be presented elsewhere; those data, however, do not alter the conclusion that ongoing selection is operating at the Mpi allozyme and is likely responsible for allelic variation among habitats and geographic populations (Table 1).

The second unknown about population differences in the patterns of selection at Mpi concerns limited gene flow and local adaptation. The data presented here help resolve this issue. The F_{ST} data in Fig. 1 show that the vast majority of polymorphic loci across the genome show little differentiation between ME and RI. Coupled with the evidence that the Mpi polymorphism is indeed the same change in nucleotides, and that the value of F_{ST} for both the Mpi allozyme and the causative SNP is $<5\%$ between selectively differentiated samples (Fig. 1), a parsimonious explanation is that the variable responses of Mpi and Gpi to selection gradients in ME, RI and the Miramichi region are due to differences in local ecological and physiological stressors. Moreover, mtDNA and microsatellite data for the comparison between ME and RI indicate F_{ST} values to be at or below the median for the genome-wide average, but the genomic data show 140 genomic regions that may be due to selective differentiation between these populations (Fig. 2) (Flight et al. 2012). Figure 2 shows how one can sort out those loci that are generally under diversifying selection versus those that are experiencing unique modes of selection between pairs of populations; highly differentiated loci that are shared between pairs of populations become strong candidates for further analysis. It will interesting to apply these whole-genome scans to the differentiation across the Miramichi estuary, between high and low tidal stations on the ME and RI coasts, and between pairs of sites identified in Flowerdew's (1983) original allozyme survey. Population genomic scans of these localities should uncover SNPs in Mpi and Gpi, and should identify other loci with even stronger patterns of population-specific allele frequency differentiation, such as the contig with homology to the settlement-inducing complex in *B. amphitrite* (Dreanno et al. 2006) described above. Given the long history of ecological studies in acorn barnacles, additional genomic analysis are likely to add a lot to our understanding of how selection shapes genetic variation in natural populations.

Acknowledgments

We wish to thank Dr Eric Sanford for providing specimens for this experiment, and Dr Sarah Bray for access to her laboratory for sorting the Southwold samples. Sohini Ramachandran, Andrew Kern, and the Brown University Genomics groups were instrumental for analysis and interpretation of data. Comments from anonymous reviewers provided many helpful suggestions.

Funding

This work has been supported by grants from the US National Science Foundation DEB 0108500, the National Institutes of Health GM067862, and by the Rhode Island Experimental Program to Stimulate Competitive Research (EPSCoR) Graduate Research Fellowship.

References

Avise JC, Lansman RA, Shade RO. 1979. The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus Peromyscus. Genetics 92:279–95.

Bachmann K, Rheinsmith EL. 1973. Nuclear DNA amounts in pacific Crustacea. Chromosoma 43:225–36.

Barnes H. 1953. On the southern limits of the intertidal Barnacle Balanus-Balanoides. Ecology 34:429–30.

Barnes H, Barnes M. 1954. The general biology of Balanus-Balanus (L) Da Costa. Oikos 5:63–76.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS One 6:e19315.

Bazin E, Glemin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312:570–2.

Beaumont MA. 2005. Adaptation and speciation: what can Fst tell us? Trends Ecol Evol 20:435–40.

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13:969–80.

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. Proc Royal Soc Lon Ser B-Biol Sci 263:1619–26.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E, Langley CH. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol 5:e310.

Bertness MD, Gaines SD. 1993. Larval dispersal and local adaptation in Acorn Barnacles. Evolution 47:316–20.

Brind'Amour A, Bourget E, Tremblay R. 2002. Fecundity, growth rate and survivorship at the interface between two contiguous genetically distinct groups of Semibalanus balanoides. Marine Ecol Progr Ser 229:173–84.

Brown AF, Kann LM, Rand DM. 2001. Gene flow versus local adaptation in the northern acorn barnacle, Semibalanus balanoides: insights from mitochondrial DNA variation. Evolution 55:1972–9.

Dreanno C, Kirby RR, Clare AS. 2006. Locating the barnacle settlement pheromone: spatial and ontogenetic expression of the settlement-inducing protein complex of Balanus amphitrite. Proc Biol Sci 273:2721–8.

Dufresne F, Bourget E, Bernatchez L. 2002. Differential patterns of spatial divergence in microsatellite and allozyme alleles: further evidence for locus-specific selection in the acorn barnacle, Semibalanus balanoides? Mol Ecol 11:113–23.

Flight PA. 2011. Genetic signatures of natural selection and glaciation in the nearshore North Atlantic. Department of Ecology and Evolutionary Biology. Providence, Rhode Island: Brown University. p. 229.

Flight PA, O'Brien MA, Schmidt PS, Rand DM. 2012. Genetic structure and the North American postglacial expansion of the Barnacle, Semibalanus balanoides. J Heredity 103:153–65.

Flight PA, Schoepfer SD, Rand DM. 2010. Physiological stress and the fitness effects of Mpi genotypes in the acorn barnacle Semibalanus balanoides. Marine Ecol Progr Ser 404:139–49.

Flowerdew MW. 1983. Electrophoretic investigation of populations of the cirripede *Balanus balanoides* (L.) around the North Atlantic seaboard. Crustaceana 45:260–78.

Flowerdew MW, Crisp DJ. 1975. Esterase heterogeneity and an investigation into racial differences in Cirripede Balanus-Balanoides using acrylamide-gel electrophoresis. Mar Biol 33:33–9.

Flowerdew MW, Crisp DJ. 1976. Allelic esterase isozymes, their variation with season, position on shore and stage of development in Cirripede Balanus-Balanoides. Mar Biol 35:319–25.

Futschik A, Schlotterer C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–18.

Goldstein DB, Clark AG. 1995. Microsatellite variation in North American populations of Drosophila melanogaster. Nucleic Acids Res 23:3882–6.

Gregory TR. 2011. Animal genome size database. Available online at: http://www.genomesize.com.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6:e1000862.

Holm ER, Bourget E. 1994. Selection and population genetic-structure of the Barnacle Semibalanus-Balanoides in the Northwest Atlantic and Gulf of St-Lawrence. Marine Ecol Progr Ser 113:247–56.

Jost L. 2008. G(ST) and its relatives do not measure differentiation. Mol Ecol 17:4015–26.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ. 2011. Genomic differentiation between temperate and tropical Australian populations of Drosophila melanogaster. Genetics 187:245–60.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54:595–609.

Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–95.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–9.

Li RQ, Zhu HM, Ruan J, Qian WB, Fang XD, Shi ZB, Li YR, Li ST, Shan G, Kristiansen K, Li SG, Yang HM, Wang J, Wang J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–72.

Mackill DJ, Zhang Z, Redona ED, Colowit PM. 1996. Level of polymorphism and genetic mapping of AFLP markers in rice. Genome 39:969–77.

Maynard Smith J. 1998. Evolutionary genetics. Oxford, UK: Oxford University Press.

Meiklejohn CD, Montooth KL, Rand DM. 2007. Positive and negative selection on the mitochondrial genome. Trends Genet 23:259–63.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–8.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol 13:735–48.

Rand DM, Spaeth PS, Sackton TB, Schmidt PS. 2002. Ecological genetics of Mpi and Gpi polymorphisms in the Acorn Barnacle and the spatial scale of neutral and non-neutral variation. Integr Comp Biol 42:825–36.

Schlotterer C, Pemberton J. 1994. The use of microsatellites for genetic analysis of natural populations. EXS 69:203–14.

Schmidt PS. 2001. The effects of diet and physiological stress on the evolutionary dynamics of an enzyme polymorphism. Proc R Soc Lond B 268:9–14.

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012

Schmidt PS, Bertness MD, Rand DM. 2000. Environmental heterogeneity and balancing selection in the acorn barnacle Semibalanus balanoides. Proc R Soc Lond Ser B Biol Sci 267:379–84.

Schmidt PS, Rand DM. 1999. Intertidal microhabitat and selection at MPI: Interlocus contrasts in the northern acorn barnacle, Semibalanus balanoides. Evolution 53:135–46.

Schmidt PS, Rand DM. 2001. Adaptive maintenance of genetic polymorphism in an intertidal barnacle: Habitat- and life-stage-specific survivorship of Mpi genotypes. Evolution 55:1336–44.

Southward AJ, Crisp DJ. 1954. Recent changes in the distribution of the intertidal Barnacles Chthamalus-Stellatus Poli and Balanus-Balanoides L in the British-Isles. J Animal Ecol 23:163–77.

Veliz D, Bourget E, Bernatchez L. 2004. Regional variation in the spatial scale of selection at MPI* and GPI* in the acorn barnacle Semibalanus balanoides (Crustacea). J Evol Biol 17:953–66.

Veliz D, Duchesne P, Bourget E, Bernatchez L. 2006. Stable genetic polymorphism in heterogeneous environments: balance between asymmetrical dispersal and selection in the acorn barnacle. J Evol Biol 19:589–99.

Vitalis R, Dawson K, Boursot P, Belkhir K. 2003. DetSel 1.0: a computer program to detect markers responding to selection. J Hered 94:429–31.

Wares JP. 2010. Natural distributions of mitochondrial sequence diversity support new null hypotheses. Evolution 64:1136–42.

Wares JP, Cunningham CW. 2001. Phylogeography and historical ecology of the North Atlantic intertidal. Evolution 55:2455–69.

Wethey DS. 1983. Geographic limits and local zonation-the Barnacles Semibalanus (Balanus) and Chthamalus in New-England. Biol Bull 165:330–41.

Wethey DS. 1984. Sun and shade mediate competition in the Barnacles Chthamalus and Semibalanus-a field experiment. Biol Bull 167:176–85.

Whitlock MC. 2011. G'ST and D do not replace FST. Mol Ecol 20:1083–91.

Wondji CS, Hemingway J, Ranson H. 2007. Identification and analysis of single nucleotide polymorphisms (SNPs) in the mosquito Anopheles funestus, malaria vector. BMC Genomics 8:5.

Wood HM, Grahame JW, Humphray S, Rogers J, Butlin RK. 2008. Sequence differentiation in regions identified by a genome scan for local adaptation. Mol Ecol 17:3123–35.

Downloaded from http://icb.oxfordjournals.org/ at Brown University on August 24, 2012