Transcriptome-wide polymorphisms of red abalone (Haliotis rufescens) reveal patterns of gene flow and local adaptation

PIERRE DE WIT and STEPHEN R. PALUMBI
Department of Biology, Hopkins Marine Station, Stanford University, 120 Ocean view Blvd., Pacific Grove, CA 93950, USA

Abstract

Global climate change is projected to accelerate during the next century, altering oceanic patterns in temperature, pH and oxygen concentrations. Documenting patterns of genetic adaptation to these variables in locations that currently experience geographic variation in them is an important tool in understanding the potential for natural selection to allow populations to adapt as climate change proceeds. We sequenced the mantle transcriptome of 39 red abalone (Haliotis rufescens) individuals from three regions (Monterey Bay, Sonoma, north of Cape Mendocino) distinct in temperature, aragonite saturation, exposure to hypoxia and disease pressure along the California coast. Among 1.17 × 10^6 Single Nucleotide Polymorphisms (SNPs) identified in this study (1.37% of the transcriptome), 21 579 could be genotyped for all individuals. A principal components analysis concluded that the vast majority of SNPs show no population structure from Monterey, California to the Oregon border, in corroboration with several previous studies. In contrast, an F_\text{ST} outlier analysis indicated 691 SNPs as exhibiting significantly higher than expected differentiation (experiment-wide P < 0.05). From these, it was possible to identify 163 genes through BLAST annotation, 34 of which contained more than one outlier SNP. A large number of these genes are involved in biomineralization, energy metabolism, heat-, disease- or hypoxia-tolerance. These genes are candidate loci for spatial adaptation to geographic variation that is likely to increase in the future.

Keywords: climate change, gene flow, local adaptation, ocean acidification, population structure, RNA-Seq, Single Nucleotide Polymorphism discovery, transcriptome

Introduction

As Earth’s climate continues to change, many species are likely to experience increased environmental stress (Reid et al. 2009). In particular, for marine species, sea level rise, higher temperatures and decreasing pH are affecting coastal and open ocean ecosystems throughout the world (Fabry et al. 2008; Pörtner & Peck 2010; Booth et al. 2011; Johnson et al. 2011; Philippart et al. 2011; Wassmann et al. 2011; Wernberg et al. 2011). For example, global sea surface temperature has risen by approximately 0.7 °C in the last century and is projected to increase 2-3 °C in the next 100 years, with the strongest impacts on more polar ecosystems (Solomon et al. 2007; Lough & Hobday 2011; Wassmann et al. 2011). The average pH of the ocean surface is expected to decrease from approximately 8.1 to nearly 7.7 by the year 2100 (Feely et al. 2009) because higher concentrations of atmospheric CO_2 continues to acidify surface waters (Sabine et al. 2004; Hauri et al. 2009; Hofmann et al. 2010). Changes in ocean circulation patterns may also be changing the distributions and severity of deep-sea, low oxygen zones (Stramma et al. 2010), which in combination with regional upwelling can contribute to hypoxic conditions in surface waters (Bograd et al. 2008).

The responses of marine organisms to these changing conditions are varied and complex. Transient high
temperatures can cause coral bleaching (Anthony et al. 2011), death of intertidal organisms (Wethey et al. 2011) or biogeographic shifts in fish and other marine species (Jones et al. 2010; Pörtner & Peck 2010; Booth et al. 2011; Johnson et al. 2011; Somero 2012). Increasing temperature is also often associated with increased exposure to microbes and incidence of disease, such as the withering-foot syndrome in California abalone (Moore et al. 2002, 2011; Raimondi et al. 2002; Rogers-Bennett et al. 2010). Although not all organisms are negatively affected by decreasing pH (Kroeker et al. 2010), species of calcifying organisms have often been shown to experience strong fitness effects, because increased acidity (lower pH) makes it more energy-consuming to precipitate CaCO$_3$ into shell or skeleton matrices (Feely et al. 2004; Hofmann et al. 2010; Erez et al. 2011; Stumpp et al. 2011; Yu et al. 2011). Shelled molluscs seem particularly susceptible: a large body of evidence suggests a deleterious effect of decreasing pH (Miller et al. 2009a; Tal mage & Gobler 2009; Beniash et al. 2010; Zippay & Hofmann 2010; Crim et al. 2011), possibly due to the fact that almost all mollusc shells to some extent consist of aragonite (Furuhashi et al. 2010), a relatively unstable form of calcium carbonate.

Although there has been much focus on possible future changes across the globe for pH, temperature, hypoxia and other environmental factors (Solomon et al. 2007; Bograd et al. 2008; Reid et al. 2009), it is difficult to predict how organisms will react. However, there are places, such as upwelling systems, that currently experience large environmental variations in all of the above-mentioned variables on small geographic or temporal scales (Aristegui et al. 2009; Belkin et al. 2009; Fransson et al. 2009; Hauri et al. 2009; Zhang et al. 2010). These systems might provide indicators for possible future genetic impacts of climate change.

The geographic variation in basic ocean environment may exert strong selection on populations to adapt to an environmental average or may allow local adaptation of populations to particular local conditions (Hereford 2009; North et al. 2011; Sanford & Kelly 2011). The effect of high dispersal was thought to be enough to keep a population genetically homogeneous along a range of environmental conditions (Díaz-Viloria et al. 2009; Hellberg 2009); however, population genetic theory shows that local adaptation depends on a balance of selection and gene flow (Sotka & Palumbi 2006). Recent studies of marine species at the whole-genome level increasingly demonstrate cases of particularly strong differentiation of some loci in the face of gene flow (Conover et al. 2006; Nielsen et al. 2009; Pespeni & Palumbi 2010). For example, Pespeni et al. (in prep) recently observed that there was enough standing genetic variation in natural populations of the purple sea urchin such that experimental exposure to high acidity conditions led to marked gene frequency evolution. Such whole-genome approaches are well suited to understanding the relationship of coastal environmental mosaics to potential adaptation for climate change effects.

Along the upwelling-dominated California coast, strong variation in temperature (Checkley & Barth 2009), oxygen concentration (Bograd et al. 2008; Connolly et al. 2010) and pH (Feely et al. 2008; Hauri et al. 2009) suggests that there may be adaptive genes in species that live across this range associated with tolerance to these stressors (Somero 2010; Pespeni et al. 2011). Abalone (*Haliotis spp.*) are semi-sessile marine gastropods that have evolved and diversified along the Pacific coast of North America (Estes et al. 2005). All but one species is currently listed as under concern in California (CADFg 2011), and they have been shown to be negatively impacted both by increasing temperatures causing increased disease occurrence (Rogers-Bennett et al. 2010) and by decreasing pH causing abnormal larval development in an experimental setting (Crim et al. 2011). The red abalone (*Haliotis rufescens*), the only species for which there is still a limited recreational fishery, shows remarkably little population structure over large geographic distances, despite a low potential for gene flow (Burton & Tegner 2000; Grunenthal et al. 2007). In this study, we hypothesize that within an overall homogeneous genetic structure in *H. rufescens*, genes that are acted upon by local natural selection may diverge between areas of different environmental stressors along the California coast. As a result, we predict that genes related to shell biomineralization, resistance to hypoxia, temperature tolerance and resistance to pathogens will show the strongest patterns of local adaptation. To test this, we sequenced the mantle transcriptome of 39 *H. rufescens* individuals from three different regions in the northern part of the species’ range between Monterey and the Oregon border (ca. 600 km) to compare Single Nucleotide Polymorphisms (SNPs) among locations. We scored all individuals at 21 579 loci and analysed them for geographic patterns and statistical outliers. We used these analyses to simultaneously estimate overall patterns of gene flow and to trace possible loci and gene functions under local environmental selection.

**Material and methods**

Non-lethal mantle tissue samples of *Haliotis rufescens* were taken *in situ* with biopsy forceps from 39 individuals in the intertidal zone or underwater using SCUBA. Sampling sites were as follows: Hopkins Marine Station (H) (*n* = 13), 36° 37.244′N, 121° 54.118′W; Fort Ross (FR)
very strong during the late summer months, a more Oregon border, has year-round upwelling, which is (L. Rogers-Bennett, personal communication). Finally, the presence of the agent of the withering-foot syndrome (et al. 2009) and only recent reports of the withering-foot syndrome during warm periods (Moore et al. 2002). Sonoma, with stronger upwelling during the summer (although primarily offshore), has temperatures between 9 and 12 °C (National Buoy Data Center, NOAA), a reported decrease during summer in aragonite saturation state (although still supersaturated in surface water) (Hauri et al. 2009) and only recent reports of the presence of the agent of the withering-foot syndrome (L. Rogers-Bennett, personal communication). Finally, the region located north of Cape Mendocino, up to the Oregon border, has year-round upwelling, which is very strong during the late summer months, a more variable temperature (between 8 and 15 °C) (National Buoy Data Center, NOAA), reported aragonite undersaturation in surface waters (Feely et al. 2008) but no presence of the agent of the withering-foot syndrome (Moore et al. 2002; L. Rogers-Bennett, personal communication).

Tissue samples were allowed to soak in RNA later (Qiagen, Valencia, CA, USA) at room temperature or at 4 °C for about 1 day, then stored at −80 °C. Total RNA was extracted from tissue using TriZol reagent (Invitrogen, Carlsbad, CA, USA). Tissue pieces were first cut into smaller pieces using a razorblade, after which they were flash frozen with liquid nitrogen and powdered with a micropestle before adding TriZol. Each sample was extracted in duplicate. A standard extraction protocol was used (Rio et al. 2010), at the end of which the RNA pellets were eluted in 25 μL of RNase-free water; duplicates were combined for a total of 50 μL. Total RNA was quantified using QuBit 2.0 fluorometer (Invitrogen) RNA assays.

For cDNA library construction, Illumina’s TruSeq RNA sample prep kit (Illumina, San Diego, CA, USA) was used, following the standard protocol, using the full 50 μl of total RNA extract as input. No optional in-line controls were used. For multiplexing, we used six different indexing adapters. Complementary DNA was quantified using a QuBit 2.0 fluorometer (Invitrogen), an Agilent Bioanalyser (Agilent, Santa Clara, CA, USA) as well as quantitative PCR, after which separate libraries were pooled into seven tubes [six with six samples each, and one with three samples (AB51, GC28 and H4)]. Cluster generation and parallel sequencing of single-end 50 base-reads was then performed in seven lanes in an Illumina HiSeq 2000 (Illumina) at the University of Utah’s Bioinformatics Core Facility.

A detailed pipeline of all data processing steps can be found at http://sfg.stanford.edu (De Wit et al. 2012). Short reads were trimmed using a phred-scale quality score cut-off of 20; residual adapter sequences were clipped from the reads after which reads shorter than 20 bases were discarded. Quality score and nucleotide distributions as well as the proportions of duplicate reads were examined using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Trimmed and clipped short reads from all individuals were imported to CLC genomics workbench (CLC bio, Cambridge, MA, USA) and combined to generate a de novo assembly with mismatch cost 1 and insertion and deletion cost 2 (limit 5), while voting for conflict resolution and ignoring all nonspecific matches and specifying a minimum contig length of 200 bases.

To annotate the de novo assembly, it was compared against the UniProt knowledgebase and GenBank’s nr database using the BLASTx algorithm, and against Gen-

Fig. 1 Map of the species range of red abalone along the US West Coast, showing collecting locations. Location codes: CC, Crescent City; T, Trinidad Harbour; PP, Patrick’s Point; AB, Anchor Bay; GC, Gerstle Cove; FR, Fort Ross; SP, Salt Point; H, Hopkins Marine Station.
Bank's nt database and the Lottia gigantea genome project's (http://genome.jgi.doe.gov/Lotgi1/Lotgi1.home.html) 'all gene models' data set using the tBLASTx algorithm. In all cases, NCBI's default parameters were used. Hits with $e$-values $<10^{-5}$ were considered significant. BLASTing was conducted on Stanford University's Bio-X2 cluster (http://bio2.stanford.edu/). Gene Ontology (GO) annotation was acquired from AGbase (http://www.agbase.msstate.edu/) using UniProt identifiers, when possible. To remove non-abalone contigs, MEGAN 4 (http://ab.inf.uni-tuebingen.de/software/megan/) (Huson et al. 2011) was used with the results of the tBLASTx to nt, to create a list of contigs that were assigned to non-metazoan (minimum score 55, minimum support 2). These contigs were excluded from further analysis.

Samples were then individually mapped against the assembly using the Burrows–Wheeler Aligner (http://bio-bwa.sourceforge.net/) (Li & Durbin 2009), using default parameters (except $n = 0.005$ and $k = 5$). For SNP detection, the Genome Analysis Toolkit (McKenna et al. 2010; DePristo et al. 2011) was used, following the best practise protocol version 2 of the Broad Institute (http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v2); only deviating in the omission of the Base Quality Score Recalibration step, as this step requires known variant sites as input. An initial variant call set was obtained using a phred-scale quality threshold of 30 as well as applying hard filtering with all settings as recommended by the Broad Institute except cluster-window size = 10. This call set was used to train the Variant Quality Score Recalibration model and filter a call set with a variant quality score threshold of 4 (all other settings as recommended by the Broad Institute). The HARD_TO_VALIDATE filter was used to train the Gaussian mixture model but was ignored when applying the final filter, as recommended by the Broad Institute. The list of polymorphic sites was then filtered to extract only sites for which there were confident genotypes (phred-scale threshold 20) assigned to all 39 individuals.

For data analysis purposes, the individuals were grouped into three populations: Monterey ($n = 13$), Sonoma (including FR, GC, SP and AB) ($n = 14$) and north of Cape Mendocino (including T, PP and CC) ($n = 12$). A principal components analysis (PCA) was performed using the EIGENSOFT software package (http://genepath.med.harvard.edu/~reich/Software.htm) (Patterson et al. 2006). The results of the PCA led us to remove two deviating individuals from Patrick's Point (PP1 and PP3) and then repeat the SNP detection pipeline as described above with the remaining 37 individuals, after which another PCA was performed.

Finally, the data set was further trimmed by removing variant sites for which all individuals had been genotyped as heterozygotes (likely to be due to either contamination or paralogous genes, $n = 70$) and by removing all sites in which the minor allele frequency was $<2.5\%$. Mean global and pairwise $F_{ST}$ were calculated using GenePop (http://kimura.univ-montp2.fr/~rousset/Genepop.htm) (Rousset 2008). To further assess the species status of individuals PP1 and PP3, genomic DNA was back-extracted from Trizol, and the VERL gene, which has been shown to be useful for abalone species identification (Gruenthal & Burton 2005), was amplified using the protocol and primers described in Swanson et al. (2001).

To scan the transcriptome for $F_{ST}$ outlier SNPs, the Lositan selection workbench (http://popgen.eu/soft/lositan/) (Antao et al. 2008) was used, using a 'neutral' mean $F_{ST}$, running 500 000 simulations and a 95% confidence interval. From the resulting list of outlier SNPs, genes for which there were more than one outlier SNP as well as at least one significant BLAST hit were extracted. Pairwise $F_{ST}$ between the three populations were calculated for outlier SNPs in these contigs using GenePop and plotted in a 3D scatterplot in R.

## Results

### Transcriptomic coverage

The number of reads per individual after quality trimming and adapter clipping ranged from 22 885 713 to 25 237 104 in the lane with only three individuals, and from 9 721 592 to 21 646 749 in the remaining lanes. 355 678 562 reads were used in the de novo assembly, creating 162 928 contigs of average length 522 bases (N50 = 653) for a total length of 85 063 104 bases (supplementary material 1). BLAST annotations could be assigned to 48 004 out of 162 928 contigs ($e$-value threshold $10^{-5}$) (supplementary material 3). BLASTx against NCBI's nr database resulted in 33 049 hits, while tBLASTx against NCBI's nt database returned 42 603 hits. BLASTx against the UniProt knowledgebase returned 28 819 hits. tBLASTx against the Lottia gigantea genome resulted in 34 994 hits. The MEGAN algorithm was able to identify 546 contigs in the reference assembly that were non-metazoan (supplementary material 4).

This study provides the first transcriptomic resource for abalone. The large number of contigs (162 928) in the de novo assembly is disconcerting (although a known issue when dealing with short reads). After removing the 546 non-metazoan contigs, this is still many more than the number of genes described for Lottia gigantea (23 851) (Lotgi v1.0), or ca. 28 000 in the sea urchin Strongylocentrotus purpuratus (Sodergren et al. 2001).
weights indicates that the differences observed derive from subtle patterns at many base pair sites and that these two individuals are heterozygous at 498 of the 500 most highly weighted SNPs on PC axis one. The VERL sequences of the two individuals were compared to the data set published on GenBank by Swanson et al. (2001) for Haliotis rufescens; both were found to be identical to sequence AF250893. Still, because of the possibility that these two individuals represent hybrids (first generation or later backcross, see Discussion), this data set was discarded and the SNP detection pipeline was repeated, resulting in 1 166 275 variant sites (Q > 20) (supplementary material 2), of which 21 579 have confident genotype calls (Q > 20) for all 37 individuals. A PCA of this data set (Fig. 2B) continues to suggest that there is no strong genetic differentiation between geographic locations. In addition, mean $F_{ST}$ indicates low genetic differentiation among regions for all SNPs (Global $F_{ST}$ = 0.0003). The distribution of minor allele frequencies in the data set is given in Fig. 3.

Only 1.85% of the detected variant sites have confident genotypes for all 37 individuals. Due to randomness in sample preparation and the multiplexing of six individuals in each HiSeq lane, it is likely that only the most highly expressed genes were sequenced to sufficient depth to be confidently genotyped for all individuals. The GATK uses a combination of sequencing depth and read qualities at a given site to calculate Bayesian posterior probabilities for the different genotypes and log odds ratios between them, emitting a genotype only if the log odds ratio exceeds a threshold ($P > 0.99$ in our case). Deeper re-sequencing would undoubtedly increase genotype data significantly and provide information on less common transcripts. For this initial study, however, genotype information at 21 579 SNP sites allows us to explore patterns of differentiation across genes likely to be important in organism function.

**Outlier and gene association analysis**

From the 21 579 SNP sites for which there was genotype information for all 37 individuals, 70 were removed due to all individuals being genotyped as heterozygotes and 8 292 were removed due to a minor allele frequency <2.5%, leaving 13 217 variant sites from 3 342 contigs for the $F_{ST}$ outlier analysis. Lositan indicated 691 SNPs as significantly more divergent between geographic locations than likely from chance alone (supplementary material 5). The SNPs were distributed across 486 contigs, from which 163 genes were identifiable by BLAST search. Of these 163, 34 genes contain more than one outlier SNP and thus are particularly good candidates for being targets of differentiation caused by natural selection (Table 1). This list includes genes basic
to shell formation and Calcium regulation, genes involved in hypoxia and disease resistance, as well as genes involved in energy metabolism. We compared pairwise $F_{ST}$ among the three populations for each SNP within the 34 genes. For 18 genes where all SNPs in a gene showed a clear consistent pattern, we list which of our three main populations was deviant (Table 1).

**Discussion**

Our data broadly agree with prior results that show very low genetic differentiation of red abalone along the California coast (Burton & Tegner 2000; Gruenthal et al. 2007), despite their predicted low dispersal ability (Prince et al. 1987). Nevertheless, we also detected outlier SNPs with higher than expected differentiation in 486 contigs, suggesting that overlain on a pattern of high gene flow is the signature of local adaptation in some genes.

**Possible hybrids**

Two individuals from Patrick’s Point (PP1 & 3) are clearly deviant in the principal component analysis.
The strongest component of this difference was derived from more than 500 loci that loaded heavily onto PCA axis 1 (Fig. 2A). Individuals PP1 and PP3 were uniquely heterozygous at all but two of these SNPs, suggesting that they might be hybrids between red and another co-occurring abalone species, such as *Haliotis walallensis* or *Haliotis kamtschakana*. Abalone hybridization is presumed rare because of the species specific interaction of gamete recognition proteins (Swanson & Vacquier 1995; Clark et al. 2009), yet natural hybrids have been reported between *Haliotis rufescens* and all other North American species of abalone at an average rate of 0.2% (de la Cruz & Gallardo-Escarate 2010). This rate was observed to be higher in areas of ecological disturbance or at the extreme range limits of one species. The collection point for PP1 and PP3 is north of Cape Mendocino, near the northern limit of *H. rufescens*. VERL sequences of the two individuals were unambiguous and identical to published sequences for *H. rufescens* (Swanson et al. 2001). The VERL primers used (Swanson et al. 2001) have previously been used to successfully amplify VERL fragments of all North American abalone species, as the polymorphism rate in VERL is very low between species (Gruenthal & Burton 2005). Still, this result cannot rule out that the individuals are hybrids or backcrosses. Confirmation of the hybrid nature of these individuals must await genomic data from the other potential parent species.

### Population differentiation

The principal component analyses (Fig. 2) and the mean $F_{ST}$ show that there is no genetic differentiation at the vast majority of SNPs between geographic locations. This is consistent with previous findings (Burton & Tegner 2000; Gruenthal et al. 2007) and indicates that red abalone larvae have dispersed among populations along the California coast, despite their short pelagic life stage. Overall $F_{ST}$ is 0.0003, suggesting high rates of historic gene flow at most SNPs.

A comparison of heterozygosity and $F_{ST}$ may help explain the conundrum between short larval duration and low differentiation. We find nearly 1.2 million high quality, variant sites in the abalone transcriptome, resulting in a polymorphism rate of about 1.37%. This is much higher than the variation described in the human genome (0.1%) (Gibbs et al. 2003), yet similarly high numbers have been reported from sea urchins (4%) (Britten et al. 1978) and oysters (1.67%) (Curole & Hedgecock 2005; Sauvage et al. 2007). The genome-wide polymorphism rate is expected to be even higher than the rate in protein-coding expressed DNA reported here. High polymorphism in abalone suggests, as in urchins and oysters, the build-up of genetic diversity in a consistently large population with a high mutation rate. Kang et al. (2010) also found high SNP heterozygosity in *Haliotis iris* in New Zealand (3.2%, $n = 2$) and multiple microsatellite studies also demonstrate high allele numbers and high heterozygosity in many species of abalone (Huang et al. 2000; Maynard et al. 2004; Tang et al. 2004; Gutierrez-Gonzalez et al. 2007; Gruenthal & Burton 2008; Shu et al. 2008; Díaz-Viloría et al. 2009; Miller et al. 2009b).

High polymorphism rate (1.37%) suggests a large population size. If the average mutation rate per base is on the order of $10^{-8} - 10^{-9}$ per generation (or about 0.2% to 0.02% mutation per million years with a 5-year
Table 1  Genes potentially under differential selective pressure along the California coast in *Haliotis rufescens*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>Contig(s) containing outlier SNPs</th>
<th>Number of outlier SNPs</th>
<th>Divergent population</th>
</tr>
</thead>
<tbody>
<tr>
<td>78 kDa glucose-regulated protein</td>
<td>Crassostrea gigas</td>
<td>Contig147173, contig148731</td>
<td>2</td>
<td>North</td>
</tr>
<tr>
<td>Actin-related protein 2/3</td>
<td>Crassostrea gigas</td>
<td>Contig83684</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Alpha-amidating enzyme precursor 2 (LPAM)</td>
<td>Lymnaea stagnalis</td>
<td>Contig89165</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Ankyrin-1</td>
<td>Pediculus humanus corporis</td>
<td>Contig24507</td>
<td>2</td>
<td>North</td>
</tr>
<tr>
<td>Arginine kinase</td>
<td>Haliotis madaka</td>
<td>Contig90217</td>
<td>3</td>
<td>Monterey</td>
</tr>
<tr>
<td>ATP synthase B chain</td>
<td>Ixodes scapularis</td>
<td>Contig104796</td>
<td>2</td>
<td>North</td>
</tr>
<tr>
<td>Calmodulin 2</td>
<td>Haliotis diversicolor</td>
<td>Contig46443</td>
<td>2</td>
<td>Hopkins</td>
</tr>
<tr>
<td>Capping protein (actin filament) muscle Z-line, beta (Capzb) Catalase</td>
<td>Haliotis discus discus</td>
<td>Contig83915</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Crumbs complex protein (Smp_130280)</td>
<td>Schistosoma mansoni</td>
<td>Contig14237</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Cytosolic malate dehydrogenase</td>
<td>Mytilus californianus</td>
<td>Contig26366</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Dermatopontin</td>
<td>Haliotis discus discus</td>
<td>Contig115444</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>EGF-like domain containing protein</td>
<td>Haliotis diversicolor</td>
<td>Contig83774</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Galectin 4-like protein transcript</td>
<td>Haliotis discus hannai</td>
<td>Contig83577</td>
<td>3</td>
<td>North</td>
</tr>
<tr>
<td>Glucose-6-phosphate 1-dehydrogenase (g6PD)</td>
<td>Crassostrea gigas</td>
<td>Contig25921</td>
<td>3</td>
<td>North</td>
</tr>
<tr>
<td>Glutamine synthetase</td>
<td>Crassostrea gigas</td>
<td>Contig90598</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Glycine transporter 2</td>
<td>Danio rerio</td>
<td>Contig102343</td>
<td>2</td>
<td>Sonoma</td>
</tr>
<tr>
<td>Hepatopancreas 5-aminolevulinate synthase (ALS)</td>
<td>Limulus polyphemus</td>
<td>Contig88857</td>
<td>3</td>
<td>North</td>
</tr>
<tr>
<td>Hillarin</td>
<td>Hirudo medicinais</td>
<td>Contig88612</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Malate dehydrogenase precursor</td>
<td>Haliotis discus discus</td>
<td>Contig83945</td>
<td>2</td>
<td>Sonoma</td>
</tr>
<tr>
<td>Molluscan defence molecule precursor (MDM)</td>
<td>Lymnaea stagnalis</td>
<td>Contig84742</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Nervous system adducin (ADD)</td>
<td>Aphysia californica</td>
<td>Contig24566</td>
<td>4</td>
<td>Sonoma</td>
</tr>
<tr>
<td>Pif177 Aragonite-binding protein</td>
<td>Pinctada fascata</td>
<td>Contig82770</td>
<td>3</td>
<td>North</td>
</tr>
<tr>
<td>Plancitoxin</td>
<td>Haliotis cracheridii</td>
<td>Contig88923</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Poly-U binding splicing factor a (puf60a)</td>
<td>Danio rerio</td>
<td>Contig24491</td>
<td>3</td>
<td>Monterey</td>
</tr>
<tr>
<td>Proline rich 15-like (PRR15L) ATPase</td>
<td>Homo sapiens</td>
<td>Contig147114</td>
<td>2</td>
<td>North</td>
</tr>
<tr>
<td>Proteasome alpha type 2</td>
<td>Haliotis discus discus</td>
<td>Contig27448</td>
<td>3</td>
<td>North</td>
</tr>
<tr>
<td>Protein disulfide isomerase 1</td>
<td>Ixodes scapularis</td>
<td>Contig16119</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RAS-related protein</td>
<td>Ixodes scapularis</td>
<td>Contig13864</td>
<td>3</td>
<td>Sonoma</td>
</tr>
<tr>
<td>Sarco/endoplasmic reticulum calcium ATPase isoform C (SERCA)</td>
<td>Pinctada fascata</td>
<td>Contig137187, contig137221</td>
<td>2</td>
<td>Sonoma</td>
</tr>
<tr>
<td>Sarcophaga 26,29 kDa proteinase</td>
<td>Periplaneta americana</td>
<td>Contig83712</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Sodium/glucose cotransporter</td>
<td>Crassostrea gigas</td>
<td>Contig73178, contig102168</td>
<td>8</td>
<td>Monterey</td>
</tr>
<tr>
<td>Spectrin alpha chain</td>
<td>Ixodes scapularis</td>
<td>Contig24945</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>TNFAIP3 interacting protein 1</td>
<td>Danio rerio</td>
<td>Contig88834</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Divergent population is noted only when pairwise $F_{ST} > 0.1$ in comparisons with both other populations in some or all of the outlier SNPs within a gene, and no conflicting pattern is seen.
generation time (Swanson et al. 2001)), then the effective population size for red abalone is approximately ($\Theta = 4N_e \mu$) 350 000–3 500 000. Populations with such a high effective size show very low genetic drift, and as a result even small amounts of gene flow can greatly reduce differentiation (Slatkin 1993). Even under an isolation-by-distance model along a one-dimensional 2000 km coastline, this large population size is predicted to have little discernable structure (e.g. $<F_{ST} = 0.005$ coastwide) with as little average dispersal as 5–15 km. 

(effective pop = 175–1750 per km; slope = 0.01/2000 km; average dispersal = $\sqrt{(1/(4*\Delta m * \text{slope}})) = 5.4$ km) (Rousset 1999), see also Pinsky et al. (2010). This relationship places red abalone in the zone described by Waves (1998), between low ecological dispersal and high evolutionary gene flow.

**Differentiation and selection**

Such high effective population sizes also allow the influence of selection to be discerned more clearly. On average, selection is more important to differentiation than is genetic drift when the selective coefficient, $s$, is greater than the inverse of the effective population size, $N_e$ (Slatkin 1993). With $N_e = 3.5 \times 10^6$, very low selective differentials can result in the build-up of differences in allele frequencies among populations.

$F_{ST}$ outlier analysis suggests that there are many SNPs putatively under differential selection. Our initial analysis suggested 691 SNPs from 486 contigs as significant outliers. Since false positives have been shown to be an issue in this type of analysis (Narum & Hess 2011), we chose to further filter loci indicated by Lositan as being under selection in order to not overstate our conclusions. By focusing on loci for which there were two or more outliers present, we concentrate on the abalone genes most likely to be under selection. We also restrict our discussion to contigs with a robust annotation. It is worth noting, however, that even a seemingly robust BLAST annotation ($e$-value $<10^{-5}$) should be regarded as tentative in poorly known organisms, such as abalone, especially when the annotation is derived from distantly related taxa. For this reason, we have chosen to focus the discussion on genes with known function in closely related organisms, such as oysters. Finally, it is important to keep in mind that even in cases when allele frequency differences in an outlier are correlated with environmental factors, causality cannot be inferred from natural populations. The strength of a transcriptomic data set of natural variation is the ability to detect correlations and generate hypotheses of causality to be subsequently tested in experimental settings.

**Biomineralization genes**

Recently, Pespeni et al. (in prep) observed that experimental treatment of sea urchin larvae with high CO$_2$ generated strong allele-dependent selection especially among biomineralization genes. Our analysis of abalone reveals several genes involved in biomineralization that are affected by differential selection along the coast. These genes are likely to be highly expressed in our red abalone transcriptome because we expressly chose shell-forming mantle tissues as our RNA source.

Pif177—an mRNA transcript coding for two proteins critical for nacre formation in the pearl oyster *Pinctada fucata* (Suzuki et al. 2009)—is differentiated northern California compared to Sonoma and Monterey. Northern California, especially the coast north of Cape Mendocino, is strongly affected by upwelling, which seasonally brings up seawater particularly high in CO$_2$ content and low in aragonite saturation state (Feely et al. 2008). Pif177 is proteolytically cleaved post-translationally into Pif80 and Pif97 in pearl oysters, where Pif80 binds aragonite, while Pif97 contains a chitin-binding domain and is thought to help anchor the aragonite onto a chitin matrix (Kröger 2009). The Pif97 contig with three outlier SNPs (contig82770) contains five SNPs in total, all coding for amino acid substitutions. Contig82770 matches amino acids 456–526 in Pif177, a stretch not known to be part of any functional domain in oysters. The functional significance of amino acid variation in red abalone Pif97 in different CO$_2$ regimes would represent interesting future work.

In the *H. rufescens* transcriptome, 11 contigs are identified by BLAST as Pif97, yet no matches are found for Pif80. A search of the *Lottia gigantea* genome also fails to find any matches to Pif80, suggesting that the nacre formation mechanism might be different in gastropods compared to oysters.

Calcium ATPase isoform C (SERCA) is a protein involved in intracellular (sarcoplasmic reticulum) Calcium homeostasis in pearl oysters (*P. fucata*) (Fan et al. 2007) and has been shown to be upregulated in urchin larvae exposed to high CO$_2$ (Stumpp et al. 2011). Also part of this gene family, the Plasma Membrane Calcium ATPase (PMCA) gene is associated with Calcium transport across cell membranes (Wang et al. 2008), and proton pumping out of the calcifying space in corals (Zoccola et al. 2004). Contigs137187 and 137221, identified as Calcium ATPases, are divergent in Sonoma compared to the other two populations. BLAST searches against the *Lottia* genome show that these two contigs are more similar to the sarcoplasmic reticulum form of Calcium ATPase than to the PMCA form.
**Hypoxia and disease stress**

Several genes with functions related to environmental stress or disease are indicated as under differential selective pressure along the California coast. Arginine kinase has been shown to be upregulated in shrimp exposed to low oxygen concentrations and is thought to facilitate oxygen re-uptake after a hypoxic event (Abe et al. 2007). In our assembly, 5 contigs match Arginine kinase from *Haliotis midae*, out of which one (contig90360) contains two linked outlier SNPs. Arginine kinase is different in Monterey compared to Sonoma and northern California (Table 1). The canyons of Monterey Bay bring deep water from the Oxygen Minimum Zone close to shore, and hypoxic water (<ca. 60 μmol/kg) has been regularly observed in shallow water (Bograd et al. 2008; Booth et al. 2010). Incidence of hypoxia at the other locations is currently unknown.

The 78-kDa glucose-regulated protein (GRP78 (different in northern California)) is a heat-shock protein, induced with heat treatment in oysters (Yokoyama et al. 2006) and implicated as having a function salvaging incorrectly folded proteins. Our data set has two contigs of this gene with three outlier SNPs.

Disease-related genes include the sodium/glucose co-transporter protein, which has been shown to be important in resistance to bacteria during warm conditions in the oyster *Crassostrea gigas* (Huvet et al. 2004) and is different in Monterey compared to Sonoma and northern California (contigs73178 and 102168, eight outliers). The Rickettsiales-like bacterium that causes withering-foot syndrome in California abalone had until recently only been documented south of the mouth of San Francisco Bay (Moore et al. 2002, L. Rogers-Bennett, personal communication).

**Metabolism**

An ATP synthase, a 5-aminolevulinate synthase (critical for erythrocyte formation in vertebrates (Ferreira & Gong 1995)) and a Proline-rich ATPase are all divergent in the northern population compared to the other two. Within contig 104796 (one of the ATP synthases mentioned above), 2 SNPs are outliers. Five of 18 total SNPs in this contig code for amino acid substitutions, out of which one (position 194) is an outlier showing a strong latitudinal clinal pattern. In experimental ocean acidification settings, sea urchin larvae have been found to up-regulate ATP synthase (Stumpf et al. 2011) and mantle tissue ATP levels in oysters have been seen to decrease significantly (Lannig et al. 2010), indicating an increased energy demand under high CO2 stress.

**Diversity, dispersal and selection in red abalone**

Simultaneous analysis of transcriptome-wide heterozygosity, population differentiation and outlier SNPs suggests a view of abalone population structure that reconciles different views of this species. Low planktonic duration suggests that short-distance dispersal is common in abalone, and other species in this genus show genetic differentiation across short spatial scales (Huang et al. 2000; Gruenthal & Burton 2008). However, high heterozygosity suggests very long-term, high population size in red abalone, similar to that of other benthic invertebrates such as sea urchins and oysters (Palumbi 1994). Under such conditions, a small amount of gene flow or small selective difference can be more powerful in generating differentiation than genetic drift. One prediction of this model of population genetics is dominance of genetic patterns by a large number of neutral alleles with no differentiation, overlain on a set of loci with greater than expected genetic variation due to selection. We observe this pattern in red abalone: 21 000 SNPs with no significant difference among populations and about 700 (ca. 3.3%) with greater than expected gene frequency shifts.

Loci indicated as experiencing differential selective pressures along the California coast are enriched for genes that affect the calcareous shell formation, across a mosaic of ocean chemistry that is known to affect calcification (Kroeker et al. 2010). As in purple sea urchins, this may allow local selection for alleles that operate better in high CO2, low pH conditions, given a high enough pre-existing genetic variability. Stressful environmental conditions are expected to increase in severity in future oceans due to build-up of anthropogenic CO2 (Sabine et al. 2004; Feely et al. 2009), so the existence of standing genetic variation for adaptation in a future acidic ocean is an important finding. The candidate genes uncovered by our analysis do not represent a comprehensive list—our analysis will have missed selection on larval genes, reproductive genes or genes with low transcript abundance in mantle tissue. The patterns we report here suggest further work on the small-scale geographic variation of variants sites in these genes, functional studies of the relationship between alleles and fitness, and direct experiments on the impact of high CO2 on gene frequencies.

**Conclusion**

By using RNA-Seq to conduct a transcriptome-wide scan of the red abalone, *Haliotis rufescens*, we can confirm low levels of genetic differentiation between geographic locations, along with a substantial number of
loci under differential selective pressure along the California coast. The combination of high heterozygosity and low differentiation of abalone despite low dispersal potential points to a widely distributed large population linked by small amounts of gene flow. Under these circumstances, the impact of selection on gene frequencies is predicted to be particularly visible. Loci that appear to be under selection are involved in biomineralization, resistance to hypoxia, response to heat/pathogens and energy metabolism. Upwelling of low pH water to the surface in northern California (Feely et al. 2008) may be responsible for some of the differences in biomineralization genes. Exposure to hypoxia (Bograd et al. 2008; Booth et al. 2010) and pathogens (Bergen & Raimondi 2001) in Monterey Bay might also be the cause of differential selective pressures.

Acknowledgements

We would like to thank James Grunden (Humboldt State University), Laura Rogers-Bennett and Jerry Kashiwada (California Department of Fish and Game), Freya Sommer (Hopkins Marine Station) and members of the recreational abalone fishing community for collecting help. Also, thank you Dan Barshis, Jason Ladner and Melissa Pespeni for help with data analysis protocol development and script writing. This project was funded by the Partnership for Interdisciplinary Study of the Coastal Oceans (PISCO) through funding from the David and Lucille Packard Foundation, the Wallenberg foundation grant MAW2009-0056, and NSF OCE 1041222. We would also like to thank NSF award CNS-0619926 for computer resources.

References


flow marine organism; the Atlantic cod (Gadus morhua). BMC Evolutionary Biology, 9, 276.


Swanson WJ, Aquadro CF, Vacquier VD (2001) Polymorphism in abalone fertilization genes is consistent with the neutral evolution of theegg’s receptor for lysin (VERL) and positive selection of sperm lysin. Molecular Biology and Evolution, 18, 376–383.


Wernberg T, Russell BD, Moore PJ et al. (2011) Impacts of climate change in a global hotspot for temperate marine


P.D.W. is responsible for project planning, tissue collection, sample preparation, data analysis and manuscript writing. S.R.P. is responsible for project planning, manuscript writing and funding for sequencing.

**Data accessibility**

Supplementary materials, including the *de novo* transcriptome assembly, a list of all detected variant sites plus genotypes for all individuals at those sites and information on all top hits from BLAST searches to NCBI’s nr, nt, the Uniprot knowledgebase, and the *Lottia gigantea* genome, as well as GO annotation are available for download at the Dryad Digital depository (doi:10.5061/dryad.85p80).

© 2012 Blackwell Publishing Ltd