PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Genome-wide polymorphisms show unexpected targets of natural selection

Melissa H. Pespeni, David A. Garfield, Mollie K. Manier and Stephen R. Palumbi

Proc. R. Soc. B 2012 **279**, doi: 10.1098/rspb.2011.1823 first published online 12 October 2011

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2012) **279**, 1412–1420
doi:10.1098/rspb.2011.1823
Published online 12 October 2011

Genome-wide polymorphisms show unexpected targets of natural selection

Melissa H. Pespeni[1,*], David A. Garfield[2], Mollie K. Manier[1,†] and Stephen R. Palumbi[1]

[1]*Department of Biology, Hopkins Marine Station, Stanford University, Oceanview Boulevard, Pacific Grove, CA 93950, USA*
[2]*Department of Biology, Duke University, Durham, NC 27708, USA*

Natural selection can act on all the expressed genes of an individual, leaving signatures of genetic differentiation or diversity at many loci across the genome. New power to assay these genome-wide effects of selection comes from associating multi-locus patterns of polymorphism with gene expression and function. Here, we performed one of the first genome-wide surveys in a marine species, comparing purple sea urchins, *Strongylocentrotus purpuratus*, from two distant locations along the species' wide latitudinal range. We examined 9112 polymorphic loci from upstream non-coding and coding regions of genes for signatures of selection with respect to gene function and tissue- and ontogenetic gene expression. We found that genetic differentiation (F_{ST}) varied significantly across functional gene classes. The strongest enrichment occurred in the upstream regions of E3 ligase genes, enzymes known to regulate protein abundance during development and environmental stress. We found enrichment for high heterozygosity in genes directly involved in immune response, particularly NALP genes, which mediate pro-inflammatory signals during bacterial infection. We also found higher heterozygosity in immune genes in the southern population, where disease incidence and pathogen diversity are greater. Similar to the major histocompatibility complex in mammals, balancing selection may enhance genetic diversity in the innate immune system genes of this invertebrate. Overall, our results show that how genome-wide polymorphism data coupled with growing databases on gene function and expression can combine to detect otherwise hidden signals of selection in natural populations.

Keywords: population genomics; natural selection; gene flow; *Strongylocentrotus purpuratus*; ubiquitin; innate immunity

1. INTRODUCTION

Deciphering the genetic basis of adaptive evolution in natural populations is a major goal of evolutionary biology [1,2], and has long used molecular tools to chart the adaptive role of candidate genes [3–8] or identify genetic targets of natural selection [2,9–11]. Recently, genome-wide studies in *Homo sapiens*, *Drosophila*, *Arabidopsis* and strongylocentrotid sea urchins have identified genes and regulatory regions that have undergone strong selection by examining evolutionary rate as a function of gene function or expression [12–15]. These studies have shown that shifts in habitat, morphology or physiology among species are paralleled by shifts in gene sequence. These studies suggest another approach to describe the genomic landscape on which natural selection acts within species: the ability to combine datasets on molecular signatures of natural selection with knowledge of gene function and the timing- and tissue-specificity of gene expression. By examining a large number of genes that are expressed at specific times, in particular tissues, or that play a particular functional role and by correlating these attributes to signatures of natural selection at individual loci, a much finer map of the influence of natural selection may be realized.

Previously, we developed a new polymorphism detection and genotyping approach [16] that quickly provides data on gene frequency differences across populations for a large number of loci. In the widely distributed purple sea urchin, *Strongylocentrotus purpuratus*, we identified 9112 polymorphic markers in non-coding upstream and coding regions of genes. Using Wright's *F*-statistic (F_{ST}) as a metric of population differentiation [17], we found that most of the variation was shared across latitudes along the west coast of North America (mean $F_{ST} = 0.029$) [16], as expected from previous studies of population genetics in this high-dispersal species [18,19]. However, the F_{ST} distribution was broader than expected from a randomized simulation and there were many loci with very high F_{ST} values [16].

These data provide an opportunity to test for signatures of natural selection across the sea urchin genome by taking advantage of detailed information about functional role and expression patterns for many of the genes for which we have polymorphism data. We start by using genetic differentiation among populations (F_{ST}) at all loci as proxies for the action of natural selection among natural populations [12–14]. Genome scans and outlier tests have been widely used to discover loci with higher than expected geographical differentiation [20–22]. Here instead, we test for relationships between high F_{ST} values

* Author for correspondence (mpespeni@stanford.edu).
† Present address: Department of Biology, 107 Life Sciences Complex, Syracuse University, Syracuse, NY 13244, USA.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2011.1823 or via http://rspb.royalsocietypublishing.org.

Received 31 August 2011
Accepted 23 September 2011

and tissue- or stage-specific gene expression. We then augment our analysis to include functional groups of loci and test for differences of F_{ST} based on gene ontology (GO) categories. If high F_{ST} values are distributed randomly across the genome, they will have no relationship to gene function, tissue expression or developmental timing. The alternative is that some suites of genes with high genetic differentiation are clustered within functional categories or at life-history stages where and when selection acts.

Using this approach, we identify two strong signatures of selection acting across the genome of the purple sea urchin: (i) high F_{ST} loci are concentrated in the upstream regions of ubiquitin-related genes and (ii) highly heterozygous loci are concentrated in immunity-related genes. These data reveal genetic patterns across the genome of an organism that has been a model for developmental biology [23,24], and take advantage of the species' distribution across a broad latitudinal distribution to map the influence of natural selection within and between populations.

2. MATERIAL AND METHODS

(a) *Study system*

The purple sea urchin lives from the cold waters of Alaska to the warmer waters of Baja California, Mexico [25]. They have a pelagic early life-history phase with feeding pluteus larvae that spend several weeks to months in the water column before metamorphosing into a juvenile urchin [26,27]. Previous genetic studies on mitochondrial and allozyme loci found little to no population structure along the species range [18,19], suggesting that gene flow is high, although some populations show distinct mitochondrial haplotype patterns along the Baja California peninsula in Mexico [28]. Their role as a model organism for developmental biology has resulted in the sequencing of their complete genome [23]. With a 800 Mb genome, the purple sea urchin genome structure is similar to that of humans with a similar number of genes, approximately 28 000, and similar intron and exon sizes [23].

(b) *Polymorphism and genotype data*

We identified polymorphisms and generated genotype data using restriction site tiling analysis (RSTA) [16]. We compared the genomes of 10 individuals from each of two distant locations along the species range, Boiler Bay, OR and San Diego, CA. We screened for polymorphisms at 50 935 loci across the genome of the purple sea urchin genome [16]. Simultaneously, we identified 12 431 polymorphisms and genotyped the 20 individuals at each polymorphic locus. In total, 6859 polymorphisms were in the coding regions of genes, whereas 2253 were within 2000 bp upstream of the start codon of genes. Genotype data were used to calculate F_{ST} and heterozygosity for each locus [16]. We found no linkage disequilibrium among high F_{ST} loci [16]. Using the binomial probability distribution [29], we found that when sampling 20 alleles, the standard deviation of allele frequency estimates ranged from 0.067 to 0.111 across the range of allele frequencies. We further confirmed that allele frequency data from 10 individuals per population were consistent with a larger sample size by sequencing an additional 10 individuals per population at a subset of six randomly selected polymorphic loci. We found allele frequencies from 10 and 20 sampled individuals were highly correlated across a wide range of heterozygosities ($r^2 = 0.925$, electronic supplementary material, figure S1).

(c) *Gene expression data*

We collected adult sea urchins from Soberanes, CA, near Garrapata State Park (latitude 36.44, longitude −121.92), common-garden acclimated them at the Hopkins Marine Station (Pacific Grove, CA, USA), and fed them kelp (*Macrocystis pyrifera*) ad libitum. For adult somatic tissues, we extracted and pooled RNA from eight individuals; four females and four males. For ovary and testis, we extracted and pooled RNA from reproductively active individuals: four females and four males, respectively, for each tissue. We raised larvae for two weeks in artificial sea water at 15°C following standard protocols [27] before RNA extraction (Qiagen RNeasy). We collected gene expression data for the 28 036 predicted genes across the purple sea urchin genome for three adult somatic tissues (tube foot, spine base muscle and coelomocytes), ovary, testis and two-week old larvae using custom high-density oligonucleotide Agilent arrays designed to the purple sea urchin in a previously published study [15]. To reduce the signal of non-specific binding, we subtracted three times the median value of the negative controls for each array, consistent with previously published methods [15,30]. We translated expression data into binary data, expressed and not expressed, in order to mask variation in magnitude of expression among genes and among tissues. We considered a gene expressed if the median value of all probes for a given gene had an expression value greater than zero following background subtraction. This resulted in genes expressed in at least one tissue or stage for approximately 50 per cent of the genes for which we had polymorphism data: 3339 genes expressed out of the 6859 polymorphisms in coding regions (average 1435 in each tissue per stage, electronic supplementary material, dataset S1) and 1154 expressed out of the 2253 polymorphisms in the upstream regions of genes (average 517 in each tissue/stage, electronic supplementary material, dataset S1).

To test for F_{ST} enrichment in tissue- and life-history-specific classes of genes, we characterized genes based on when and where they were expressed uniquely in six gene expression datasets from ovary, testis, coelomocytes, tube feet, spine base muscle and two week old larvae [15]. For example, 'testis only genes' were expressed in testis and not in any of the other five gene expression datasets. We compared F_{ST} distributions of genes uniquely expressed in specific tissues or life-history stage versus all genes that are not members of that unique category for which we had expression data. Specifically, we used an extension of the Pearson moment correlation, the point biserial correlation, to test for a relationship between a dichotomous variable, membership in a category or not, and a continuous variable, F_{ST} [12,31],

$$r_{pb} = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}},$$

where M_1 is the mean score of all the genes in the category, M_0 is the mean score of all the genes in our study not in that category and the number of polymorphisms in each group are n_1 and n_0, and s_{n-1} is the standard deviation when populations are sampled:

$$s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

r_{pb} can range in value from −1 to +1 [31,32].

(d) *Gene function analysis*

To determine if high F_{ST} polymorphisms were non-randomly distributed across gene regions, we tested for over-representation of high F_{ST} polymorphisms in each GO category [33]. We categorized each gene using UniProt identifiers [34] and GO biological process categories [33]. We tested for a correlation between membership in a particular GO category and F_{ST} value (point-biserial correlation, r_{rb}) as described above and as performed in Haygood *et al.* [12] and Oliver *et al.* [15]. For example, we compared F_{ST} values for all genes in the 'ubiquitin-dependent protein catabolic process' GO biological process category versus all those not members in that category. In this way, a test was performed for each biological process category with at least 10 members in the category. We hypothesized that independent selection pressures may act on coding and upstream non-coding gene regions as has been observed in humans [12]; we, therefore, performed analyses separately for polymorphisms in coding and upstream genomic regions. We also characterized loci based on when and where genes were expressed, in larvae or in adult somatic tissue, because larvae and adults inhabit different environments and may be subject to different environmental pressures [35,36]. We characterized a locus as 'adult' if it was in a gene expressed in any of the adult somatic tissues (tube foot, spine base and coelomocytes). We characterized a locus as 'larval' if it was expressed in the two week old larvae. This resulted in six datasets: all, expressed in adults, and expressed in larvae for both coding and upstream loci.

To test for enrichment of functional categories for highly heterozygous loci, we used the same six gene groups and point-biserial statistics as described above. Observed heterozygosity was calculated as the proportion of heterozygous individuals across all 20 individuals. We calculated q-values [37] to define the false discovery rate owing to multiple tests. We chose a q-value cutoff of 0.01, which resulted in the possibility of less than a fraction of a category as a potential false positive in each dataset (see electronic supplementary material, table S1 for a summary of the number of categories tested, the number of categories significant at $q < 0.01$, critical p-values and potential number of false positives).

Mean F_{ST} and heterozygosity values presented in tables 1 and 3 represent the mean across all polymorphisms in a category, from the high-value loci driving the significant enrichment to the low-value loci that are also members in a particular category. Mean values in a significantly enriched category are always higher than the mean value of polymorphisms not in the category, and are over an order of magnitude higher than the genome-wide mean F_{ST} of 0.029 or mean heterozygosity of 0.240.

Disease incidence driven by multiple pathogens is very high in southern populations of *S. purpuratus*, but much lower in northern populations [38]. If higher heterozygosity in immune-related genes is the result of natural selection in response to higher pathogen load, we might expect higher heterozygosity in the immune-related genes in San Diego, CA over Boiler Bay, OR. To test this hypothesis, we calculated heterozygosity for each population for the 15 polymorphisms that were represented across all of the enriched immune-related categories. We used a one-tailed Wilcoxon signed-rank test for paired data (implemented in R, Wilcox test). To correct for the significant difference in heterozygosity between populations across all loci (figure 2, *Wilcoxon* $p = 1.301 \times 10^{-11}$ and [16]), for each population, we subtracted the mean heterozygosity of all loci from the heterozygosity of each of the 15 immunity genes under consideration.

For both F_{ST} and heterozygosity analyses, we treated multiple polymorphisms in the same gene independently. However, in most cases, owing to the density of the markers, there was only one polymorphism per unique gene. We treated polymorphisms independently because mutation and recombination rates in the purple sea urchin are extremely high [23,39], such that polymorphisms in various segments of a gene may be subject to different evolutionary forces (mutation, selection and drift), resulting in different allele frequencies among populations. In addition, we found no linkage disequilibrium among the high F_{ST} loci [16]. To ensure that linked polymorphisms were not driving significance of highly enriched categories, we confirmed that significantly enriched categories did not have multiple polymorphisms in the same gene.

3. RESULTS

(a) *Relationship between F_{ST} and gene expression*

Overall, F_{ST} outliers (those with $p < 7 \times 10^{-6}$ and F_{ST} from 0.21 to 0.45) were in the coding regions of a chromatin assembly factor (SPU_026 263), a transcription factor (SPU_015 723) and a mannose receptor (SPU_000 294) [16]. Among these loci, only the chromatin assembly factor appeared in our array expression dataset, and this gene is expressed strongly in all tissues except ovary. To expand this approach beyond single loci, we tested for a correlation between unique expression in a specific tissue or life-stage and F_{ST} for each of the six gene expression datasets. We had gene expression data for about 50 per cent of the genes for which we had polymorphism data. Of these, on average, 7 and 6.4 per cent of genes were expressed in a single tissue or stage, in coding and upstream polymorphisms, respectively. We found that F_{ST} distributions across tissues were highly similar in both coding and upstream regions (figure 1*a*,*b*). Despite our expectation that genes expressed exclusively in larvae would harbour greater differentiation, we conclude from these data that F_{ST} distributions between Boiler Bay and San Diego are similar across tissue types in purple sea urchins.

(b) *Relationship between F_{ST} and gene function*

We found significant enrichment for high F_{ST} polymorphisms in the upstream regions of genes primarily in categories related to protein catabolism across all three upstream polymorphism datasets: all upstream polymorphisms, those upstream of genes expressed in adults, and those upstream of genes expressed in larvae (table 1, $q < 0.01$, electronic supplementary material, dataset S1). Loci in these enriched categories ($q < 0.01$) had a mean F_{ST} of 0.06 compared with 0.03 across all upstream polymorphisms (table 1 and electronic supplementary material, dataset S1). About half of the 47 unique genes are involved in ubiquitin-mediated proteolysis, and of these, 13 are E3 ligases, the third and final enzyme in the ubiquitin pathway (table 2). In addition, several other types of proteins involved in ubiquitin-mediated proteolysis were over-represented among high F_{ST} polymorphisms (table 2).

(c) *Regulatory motifs in upstream regions of E3 ligase genes*

The high F_{ST} polymorphisms we observe in E3 ligases occur within 2000 bp upstream of the start codon. The F_{ST} values

Table 1. Biological process categories enriched for high F_{ST} polymorphisms.

biological process category	no. of loci	mean F_{ST}	*q*-value
all upstream polymorphisms			
response to protein stimulus	10	0.068	0.00069
response to unfolded protein	10	0.068	0.00069
polymorphisms upstream of genes expressed in adults			
ubiquitin-dependent protein catabolic process	30	0.047	0.00946
proteolysis involved in cellular protein catabolic process	30	0.047	0.00946
cellular protein catabolic process	30	0.047	0.00946
modification-dependent protein catabolic process	30	0.047	0.00946
protein catabolic process	30	0.047	0.00946
modification-dependent macromolecule catabolic process	30	0.047	0.00946
polymorphisms upstream of genes expressed in larvae			
macromolecule catabolic process	12	0.066	0
cellular macromolecule catabolic process	12	0.066	0
biopolymer catabolic process	12	0.066	0
ubiquitin-dependent protein catabolic process	10	0.065	0.00001
proteolysis involved in cellular protein catabolic process	10	0.065	0.00001
protein catabolic process	10	0.065	0.00001
modification-dependent protein catabolic process	10	0.065	0.00001
cellular protein catabolic process	10	0.065	0.00001
modification-dependent macromolecule catabolic process	10	0.065	0.00001
post-translational protein modification	19	0.051	0.00276
total number of unique genes represented	47	0.059	

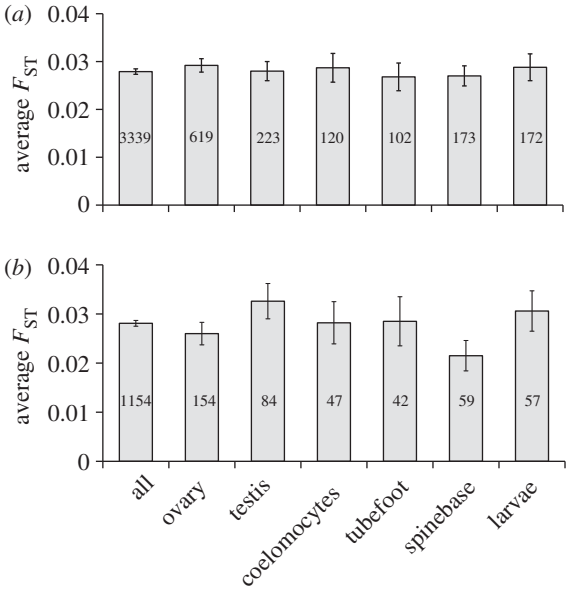

Figure 1. Average F_{ST} values for genes uniquely expressed in specific tissues and life-history stages in (*a*) coding and (*b*) upstream regions. Error bars reflect standard error. Numbers show the number of loci uniquely expressed in each category. Statistical significance was calculated using point biserial correlation between genes expressed uniquely in a tissue or stage versus the rest of expressed genes not expressed in a particular category; all tests were non-significant.

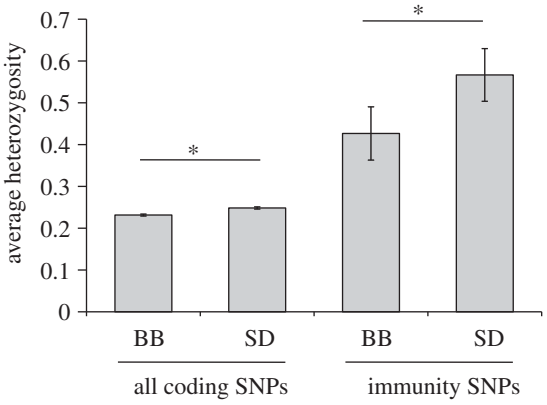

Figure 2. Average heterozygosity across all coding loci ($n = 6859$) and immunity loci present in all enriched immunity categories ($n = 15$) in Boiler Bay, OR and San Diego, CA (all coding SNPs: Wilcoxon $p = 1.301 \times 10^{-11}$, immunity SNPs: Wilcoxon $p = 0.022$ correcting for difference across all loci by subtracting the genome-wide mean heterozygosity from the heterozygosity of each immunity locus for each population). Error bars reflect standard error, asterisks represent significant differences between groups.

for polymorphisms upstream of coding regions are positively correlated ($r^2 = 0.32$, $p < 0.0001$) with F_{ST}s of polymorphisms in adjacent exons [16]. As a result, E3 ligase polymorphisms could be linked with important functional polymorphisms in the adjacent coding regions.

Alternatively, E3 ligase polymorphisms may be in regions that control E3 ligase expression. Empirical mapping of regulatory sites reveals that many important regulatory elements are located in the 5′ untranslated region and immediately upstream of the translation start site (see electronic supplementary material, text S1 and references therein). To test whether the upstream regions of these genes could be playing a regulatory role, we looked for over-represented motifs among the upstream regions of the E3 ligases using the program MEME (http://meme.nbcr.net.) with the minimum motif width set to 6 and the maximum to 14. The results were then compared with results from running MEME on shuffled versions of the same data. The results show strong evidence for three over-represented motifs that are independent of base composition (*p*-values range from 10^{-9} to 10^{-82}, see electronic supplemental material, figure S2). Shuffled data show no such over-representation. Likewise, a parallel analysis of

Table 2. Genes with high F_{ST} polymorphisms in upstream non-coding regions in significantly enriched categories. (This list represents the 18 unique genes with F_{ST} values greater than the mean across all upstream polymorphisms (0.029). These are the genes driving enrichment of proteolysis categories.)

gene annotation	gene number	E3 ligases	F_{ST}
autocrine motility factor receptor, isoform 2 (AMFR)	SPU_027 752	✓	0.30
ankyrin-1 (ANK1)	SPU_014 484	✓	0.18
ubiquitin carboxyl-terminal hydrolase 30 (USP30)	SPU_010 368	✓	0.14
Ufm1-conjugating enzyme 1 (UFC1)	SPU_017 244	✓	0.14
poly(A)-specific ribonuclease (PARN)	SPU_025 761		0.11
RING finger protein 34 (RNF34)	SPU_004 253	✓	0.09
torsin A precursor (TOR1A)	SPU_014 383		0.08
heat shock cognate 71 kDa protein (HSC70)	SPU_014 676		0.08
CREB3L1 protein (OASIS)	SPU_006 286		0.06
E3 ubiquitin protein ligase (TRAF7)	SPU_017 483	✓	0.06
midline-1 (MID1)	SPU_017 855	✓	0.05
heat shock cognate 71 kDa protein (HSC70)	SPU_009 479		0.05
HECT domain and RCC1-like domain protein 3 (HERC3)	SPU_003 837	✓	0.05
ankyrin repeat and SOCS box protein 11 (ASB11)	SPU_028 725	✓	0.05
ZYG11B protein	SPU_022 191	✓	0.05
ankyrin repeat and SOCS box protein 13 (ASB13)	SPU_023 367	✓	0.05
midline-1 (MID1)	SPU_007 506	✓	0.03
ankyrin-1 (ANK1)	SPU_007 299	✓	0.03

upstream regions of eight sea urchin genes for which we have a large amount of information on regulatory function (Sp-AN; CyIIa; CyIIIa; FoxB; HE; SM30-beta; SM50; Endo16, see references in electronic supplementary material, text S1) shows a similar number and density of over-represented motifs.

(d) *Relationship between heterozygosity and gene function*

There were no strong patterns of enrichment for high heterozygosity in the upstream regions of genes or in genes expressed in larvae or adult somatic tissues after multiple test correction (electronic supplementary material, dataset S1). Heterozygosity was significantly enriched however, in the coding regions of genes, particularly in genes related to immune response (table 3 and electronic supplementary material, dataset S1). Forty-three of the 48 categories enriched for high heterozygosity at the $q < 0.01$ level were immune-related. The mean heterozygosity for these categories was 0.44 compared with 0.24 across all coding polymorphisms (table 3 and electronic supplementary material, dataset S1). The majority of the highly heterozygous polymorphisms driving this pattern were in five different NALP proteins (**NA**CHT-, **L**RR- and **P**YD-containing protein). NALP proteins are activated to form the 'inflammasome,' a protein complex responsible for mediating activation of pro-inflammatory caspases by Toll-like receptors during a cell's response to microbial infection [40].

To confirm that enrichment within a gene family, the NALP proteins, was not owing to cross-reactivity of DNA among RSTA probes, we attempted to align all RSTA probe sequences from NALP genes. We found that none of the probe sequences aligned despite allowing low sequence similarity, low penalties for gaps and minimal overlap (10 bases). This confirms our screen for each RSTA probe sequence to be unique, non-overlapping and match only one place in the genome [16]. We further showed that hybridization was reduced to background levels with four or more mutations in a 50 bp sequence [16].

In accord with our hypothesis, we found that heterozygosity was significantly higher in San Diego than in Boiler Bay, in immunity-related genes (figure 2, *Wilcoxon* $V = 24$, $p = 0.022$). These results suggest that natural selection may act to increase heterozygosity in immune-related genes within a population in response to pathogen load and diversity.

4. DISCUSSION

We examined genome-wide data on genetic polymorphisms along with a complementary dataset on gene expression to test for patterns of geographical differentiation of genes across tissues and life-history stages. Mean F_{ST} and the distribution of F_{ST} outliers among loci were similar among tissues and between larvae and adults. However, by collating loci into functional groups, we identified two categories of genomic loci that may be responding to natural selection across the strong environmental gradient of the purple sea urchin species range: (i) upstream regions of proteolysis genes, specifically ubiquitin-related E3 ligases and (ii) immunity-related genes, specifically NALP innate immunity, pro-inflammatory genes.

Like many highly dispersing marine species, the purple sea urchin has previously been shown to have little to no geographical population structure in neutral genes along the species range [18,19]. Our genome scan results show that indeed allele frequencies are similar across 1700 km, from Boiler Bay, OR to San Diego, CA [16], however in this study, we show that higher F_{ST} loci are concentrated non-randomly in a small set of specific functional categories. Though individual loci in these categories do not appear as outliers, the group of loci has a much higher F_{ST} than typical: over 0.06 versus 0.03, the genome-wide mean.

Such selection would usually be considered to be an example of local adaptation and might build up generation by generation to produce geographical shifts in allele frequencies. Such shifts are typical of classic cases of local adaptation at the gene level [3–8]. However, in

Table 3. Biological process categories enriched for high heterozygosity.

biological process category	no. of loci	mean H_T	*q*-value
positive regulation of inflammatory response	11	0.54	0
regulation of interleukin-18 production	10	0.54	0
negative regulation of I-kappaB kinase/NF-kappaB cascade	10	0.54	0
regulation of interleukin-18 biosynthetic process	10	0.54	0
release of cytoplasmic sequestered NF-kappaB	10	0.54	0
negative regulation of Toll signalling pathway	10	0.54	0
regulation of Toll signalling pathway	10	0.54	0
negative regulation of interleukin-1 secretion	10	0.54	0
negative regulation of cytokine secretion	10	0.54	0
negative regulation of interleukin-6 biosynthetic process	11	0.50	0
positive regulation of defence response	15	0.50	0
regulation of protein amino acid autophosphorylation	11	0.50	0
negative regulation of phosphorus metabolic process	11	0.50	0
negative regulation of phosphate metabolic process	11	0.50	0
negative regulation of protein amino acid phosphorylation	11	0.50	0
negative regulation of protein amino acid autophosphorylation	11	0.50	0
negative regulation of phosphorylation	11	0.50	0
negative regulation of cytokine biosynthetic process	12	0.48	0
positive regulation of NF-kappaB import into nucleus	12	0.48	0
positive regulation of response to external stimulus	15	0.46	0
regulation of interleukin-6 biosynthetic process	15	0.46	0
response to temperature stimulus	13	0.45	0
positive regulation of nucleocytoplasmic transport	14	0.44	0
positive regulation of protein import into nucleus	14	0.44	0
positive regulation of intracellular transport	14	0.44	0
positive regulation of transcription factor import into nucleus	14	0.44	0
negative regulation of protein kinase cascade	19	0.41	0.00004
regulation of protein amino acid phosphorylation	19	0.41	0.00004
negative regulation of secretion	13	0.43	0.00005
negative regulation of protein secretion	13	0.43	0.00005
regulation of defence response	22	0.39	0.00023
regulation of interleukin-6 production	18	0.40	0.00038
regulation of interleukin-1 beta secretion	22	0.38	0.00040
regulation of cytokine secretion	22	0.38	0.00040
regulation of interleukin-1 secretion	22	0.38	0.00040
positive regulation of cytokine secretion	22	0.38	0.00040
positive regulation of interleukin-1 secretion	22	0.38	0.00040
positive regulation of interleukin-1 beta secretion	22	0.38	0.00040
negative regulation of protein modification process	15	0.41	0.00040
regulation of inflammatory response	18	0.39	0.00108
regulation of phosphorus metabolic process	32	0.36	0.00140
regulation of phosphate metabolic process	32	0.36	0.00140
establishment or maintenance of epithelial cell apical/basal polarity	42	0.34	0.00282
establishment or maintenance of apical/basal cell polarity	42	0.34	0.00282
translational elongation	12	0.40	0.00495
negative regulation of signal transduction	51	0.33	0.00523
mRNA catabolic process	11	0.40	0.00837
regulation of NF-kappaB import into nucleus	19	0.37	0.00896
total number of unique genes represented	163	0.44	

high-dispersal species such as sea urchins, a second demographic possibility must be considered in which differential survival of dispersing larvae occurs every generation [3,41]. This type of selection, called phenotype–environment mismatch [36], has been seen for the Got-2 allozyme locus in purple sea urchins, where recruits and adults showed different allele frequencies [19]. Differences in allele frequency on the order of what we observe in this study could be owing to selection occurring every generation, though some degree of local retention of locally produced alleles may also occur in this system.

Finding these potential targets of selection in proteolysis and immunity pathways, despite high-neutral gene flow, illustrates the strength of developing and applying approaches that yield genetic data at many loci across the genome of an organism [16] and the analytical power of combining these data with genome-wide gene function and expression data. Comparing sets of genes in similar functional categories allows for a more comprehensive examination of spatial genetic differentiation than do outlier analyses of genome scans.

(a) *Adaptive evolution in the purple sea urchin: ubiquitin and protein degradation*

We found that the majority of functional categories enriched for high F_{ST} were involved in the ubiquitin–proteasome

pathway. As the principal mechanism for protein degradation in the cell, the ubiquitin-proteasome pathway plays an important role in regulating numerous cellular processes including the response to environmental stress, regulation of protein half-life, immune regulation, apoptosis, cell cycle control, DNA transcription and repair, and differentiation and development [42].

Ubiquitin proteins were demonstrated to be subject to strong positive selection in nematodes and plants [43], and differences in gene expression in four ubiquitin-related genes among four natural isolates of wine yeast were suggested to be potentially adaptive [44]. E3 ligases have specific recognition domains that recognize and bind to their target proteins. This specificity allows the addition of a ubiquitin protein tag to a particular target, thus marking it for cellular degradation. Such selective degradation allows E3 ligases to play a role in functional gene regulation through selective degradation of proteins [20]. This type of regulation is not visible in surveys of mRNA levels or patterns of gene expression, but may be a common mechanism for control of rapid development or reaction to environmental change. Because they are target-specific proteins, there are many more E3 ligases than there are other types of proteins in the ubiquitin pathway [42], perhaps accounting for their dominance of this functional category.

Regions upstream of the E3 ligases in our dataset—where the putative SNPs under selection reside—are densely over-represented with motifs highly similar to those of known regulatory regions. However, this does not prove regulatory function of upstream regions in E3 ligases nor do results prove that the RSTA polymorphisms in upstream regions alter protein regulation. Studying the environmental regulation of protein ubiquitination may be a valuable next step. The role of E3 ligases in sea urchin development, and their potential role in environmentally dependent selection during larval or adult phases would be a novel link in the chain of evidence about the role of selection across the genome. It should also be noted that, owing to sampling only two populations and the high degree of environmental heterogeneity along the species range, selection in proteolysis genes could also be driven by a number of factors besides temperature (e.g. upwelling, pH, salinity, wave action and dissolved oxygen).

(b) *Immune genes*

We found that the majority of functional categories enriched for high heterozygosity were related to immune response. Purple sea urchin populations are known to experience strong disease pressures [25,38,45,46], and have an advanced innate immune system with a large suite of Toll-like receptors (222 versus approx. 20 found in other genomes) that play a crucial role in activating an immune response to microbial infection [47,48]. Higher diversity in NALP proteins may interact well with the high diversity of Toll-like receptors and could play an important role in the ability of an individual to respond to diverse microbial stressors. High heterozygosity has been known to be an adaptive benefit maintained by balancing selection in several vertebrate immune-related genes, particularly the major histocompatibility complex in humans and natural populations of stickleback fish [49–51]. In accord with our results, previous studies in *S. purpuratus* have shown high levels of genetic diversity in a new family of innate immune genes, the 185/333 family, expressed in response to lipopolysaccharide immune challenge [52]. Most of the high heterozygosity we document is in NALP proteins that respond to Toll-like receptors and activate cellular immune response. We also show that heterozygosity is higher in San Diego than Boiler Bay populations. Higher heterozygosity in the southern part of the species range could be owing to higher rate of infection in warmer waters, higher microbial load in the water owing to proximity to higher human population density, or greater alterations of ecosystems by removal of multiple natural urchin predators [46,53]. These results suggest that there may be a relationship between disease prevalence and heterozygosity among invertebrate immune-related genes. Exploring the fine-scale population genetics of these genes across multiple populations with a range of disease incidence could test this further.

(c) *Genes expressed in different tissues or early life-history stages*

We originally hypothesized that genes expressed in larvae or different adult tissues may be subject to different selection, because in purple sea urchins larvae and adults inhabit different environments, and because different tissues play different physiological roles. Such a selective signal was observed in a previous comparison across 19 000 loci of purple sea urchins with a deep-water congener [15]. However, our test of this hypothesis turned up no striking differences. Neutral genetic differentiation among loci may predominate in these data, and any tissue-specific selection patterns within species may not be different enough to create a selective signal in this neutral noise. In this study, negative results comparing tissues and life stages may also suffer from low numbers of loci that are expressed in only one tissue. Only in ovary were there a large number of tissue-specific loci ($n = 619$), and so patterns of selection in different tissues may be spread among loci that are expressed widely. Also we had gene expression data for about 50 per cent of the genes for which we had F_{ST} data. Genes missing from this microarray dataset could be those expressed at low levels, those expressed in tissues or stages not surveyed, or those expressed in response to diverse environmental stimuli. A more complete inventory of tissue- and stage-specific genes may refine our view of selection in different parts of an organism's phenotype and ontogeny.

(d) *Outliers*

Recently, a number of powerful tools have been developed to test among loci for those that have higher F_{ST} than expected by chance alone [20–22]. These outlier tests compare a single F_{ST} value with those from across a wide range of loci. Our dataset reveals several such global outliers, with levels of genetic differentiation an order of magnitude higher than the mean [16]. However, the number of these single-locus outliers is small compared with the number of loci tested. Here, we extend this approach to multi-locus comparisons by testing sets of loci grouped by expression pattern and physiological role. These tests are similar to those that look for patterns of gene expression or molecular evolution as a function of

GO [12,13,15], but here we use F_{ST} or heterozygosity as a proxy for selection. Testing for the concentration of high F_{ST} or heterozygosity in functional categories of genes takes a more comprehensive look across the genome for signals of selection. Genes, pathways and processes implicated are candidates for further investigation.

(e) *Coding versus regulatory evolution*

We find that the coding regions of immunity genes are enriched for high heterozygosity, while the upstream putative regulatory regions of ubiquitin-related genes are enriched for high F_{ST}. These results suggest that different and independent selection pressures are acting on coding versus regulatory regions of the genome dependent on specific gene function. The distinct contributions of coding and regulatory mutations to phenotypic evolution have long been hypothesized [54,55], yet both have rarely been systematically studied in a single species or clade [56,57]. The only other genome-wide study, we know of, that scans for selection in both coding and non-coding regions [12] finds similar results in humans. In fact, similar to this study, Haygood *et al.* [12] found that the coding regions of immunity-related functional categories are enriched for positive selection. Broadly, this study is among the first outside humans to show that there are probable independent evolutionary forces acting on coding and upstream non-coding regions of the genome.

5. CONCLUSION

In conclusion, we identify signals of natural selection acting in specific regions of the genome of the purple sea urchin that may represent novel mechanisms of adaptive evolution along the strong latitudinal gradient of the west coast of North America. We integrate genome-wide genetic diversity, gene function and gene expression data to reveal that: (i) functional enrichment for high F_{ST} polymorphisms is dominated by categories related to proteolysis and (ii) functional enrichment of high heterozygosity genes is dominated by genes related to immune response. These results suggest that directional selection may act on distinct polymorphisms among distant populations in proteolysis genes despite the homogenizing effects of gene flow. Meanwhile, balancing selection across the species range could act to maximize heterozygosity in the coding regions of immune-related genes, particularly in areas of higher disease pressure. These findings illustrate the power of combining genome-wide datasets on genetic diversity and gene function to identify novel mechanisms of adaptive evolution in a high gene flow species with a broad latitudinal distribution.

This work was supported by NSF SGER-0714997 (S.R.P.), NSF Graduate Research Fellowship (M.H.P.) and the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO).

REFERENCES

1 Feder, M. E. & Mitchell-Olds, T. 2003 Evolutionary and ecological functional genomics. *Nat. Rev. Genet.* **4**, 651–657. (doi:10.1038/nrg1128)
2 Stinchcombe, J. R. & Hoekstra, H. E. 2007 Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**, 158–170. (doi:10.1038/sj.hdy.6800937)
3 Koehn, R. K., Newell, R. I. E. & Immermann, F. 1980 Maintenance of an Aminopeptidase allele frequency cline by natural selection. *Proc. Natl Acad. Sci. USA* **77**, 5385–5389. (doi:10.1073/pnas.77.9.5385)
4 Oakeshott, J., Gibson, J., Anderson, P., Knibb, W., Anderson, D. & Chambers, G. 1982 Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on different continents. *Evolution* **36**, 86–96. (doi:10.2307/2407970)
5 Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. 2006 A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**, 101–104. (doi:10.1126/science.1126121)
6 Rosenblum, E., Hoekstra, H. & Nachman, M. 2004 Adaptive reptile color variation and the evolution of the MCIR gene. *Evolution* **58**, 1794–1808.
7 Storz, J., Sabatino, S., Hoffmann, F., Gering, E., Moriyama, H., Ferrand, N., Monteiro, B. & Nachman, M. 2007 The molecular basis of high-altitude adaptation in deer mice. *PLoS Genet.* **3**, e45. (doi:10.1371/journal.pgen.0030045)
8 Jessen, T., Weber, R., Fermi, G., Tame, J. & Braunitzer, G. 1991 Adaptation of bird hemoglobins to high altitudes: demonstration of molecular mechanism by protein engineering. *Proc. Natl Acad. Sci. USA* **88**, 6519–6522. (doi:10.1073/pnas.88.15.6519)
9 Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. 2003 The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994. (doi:10.1038/nrg1226)
10 Shapiro, M., Marks, M., Peichel, C., Blackman, B., Nereng, K., Jónsson, B., Schluter, D. & Kingsley, D. 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723. (doi:10.1038/nature02415)
11 Chan, Y. *et al.* 2010 Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302. (doi:10.1126/science.1182213)
12 Haygood, R., Babbitt, C., Fedrigo, O. & Wray, G. 2010 Contrasts between adaptive coding and noncoding changes during human evolution. *Proc. Natl Acad. Sci. USA* **107**, 7853–7857. (doi:10.1073/pnas.0911249107)
13 Holloway, A., Lawniczak, M., Mezey, J., Begun, D. & Jones, C. 2007 Adaptive gene expression divergence inferred from population genomics. *PLoS Genet.* **3**, e187 (doi:10.1371/journal.pgen.0030187)
14 Turner, T., Bourne, E., Von Wettberg, E., Hu, T. & Nuzhdin, S. 2010 Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nat. Genet.* **42**, 260–263. (doi:10.1038/ng.515)
15 Oliver, T., Garfield, D., Manier, M., Haygood, R., Wray, G. & Palumbi, S. 2010 Whole-genome positive selection and habitat-driven evolution in a shallow and a deep-sea urchin. *Genome Biol. Evol* **2**, 800–814. (doi:10.1093/gbe/evq063).
16 Pespeni, M., Oliver, T., Manier, M. & Palumbi, S. 2010 Restriction site tiling analysis: accurate discovery and quantitative genotyping of genome-wide polymorphisms using nucleotide arrays. *Genome Biol.* **11**, R44.
17 Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 159.
18 Palumbi, S. R. & Wilson, A. C. 1990 Mitochondrial DNA diversity in the sea urchins *Strongylocentrotus purpuratus* and *S. Droebachiensis*. *Evolution* **44**, 403–415. (doi:10.2307/2409417)
19 Edmands, S., Moberg, P. E. & Burton, R. S. 1996 Allozyme and mitochondrial DNA evidence of population subdivision in the purple sea urchin *Strongylocentrotus purpuratus*. *Mar. Biol.* **126**, 443–450. (doi:10.1007/BF00354626)

20 Beaumont, M. A. & Balding, D. J. 2004 Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13**, 969–980. (doi:10.1111/j.1365-294X.2004.02125.x)

21 Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A. & Luikart, G. 2008 LOSITAN: a workbench to detect molecular adaptation based on a F_{ST}-outlier method. *BMC Bioinf.* **9**, 323. (doi:10.1186/1471-2105-9-323)

22 Foll, M. & Gaggiotti, O. 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993. (doi:10.1534/genetics.108.092221)

23 Sea Urchin Genome Sequencing Consortium 2006 The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952. (doi:10.1126/science.1133609)

24 Davidson, E. 2006 *The regulatory genome: gene regulatory networks in development and evolution*. New York, NY: Academic Press.

25 Rogers-Bennett, L. 2007 The ecology of *Strongylocentrotus franciscanus* and *Strongylocentrotus purpuratus*. In *Edible sea urchins: biology and ecology* (ed. J. M. Lawrence), pp. 393–425. Amsterdam, The Netherlands: Elsevier.

26 Strathmann, R. 1978 Length of pelagic period in echinoderms with feeding larvae from the Northeast Pacific. *J. Exp. Mar. Biol. Ecol.* **34**, 23–28. (doi:10.1016/0022-0981(78)90054-0)

27 Strathmann, M. F. 1987 Phylum Echinodermata, Class Echinoidea. In *Reproduction and development of marine invertebrates of the northern Pacific coast: data and methods for the study of eggs, embryos, and larvae* (ed. M. F. Strathmann), pp. 511–534. Seattle, WA: University of Washington Press.

28 Olivares-Banuelos, N. C., Enriquez-Paredes, L. M., Ladah, L. B. & De La Rosa-Velez, J. 2008 Population structure of purple sea urchin *Strongylocentrotus purpuratus* along the Baja California peninsula. *Fish. Sci.* **74**, 804–812. (doi:10.1111/j.1444-2906.2008.01592.x)

29 Sokal, R. & Rohlf, F. 1995 The binomial distribution. In *Biometry* (eds R. Sokal & F. Rohlf), pp. 71–81, 3rd edn. New York, NY: W. H. Freman and company.

30 Wei, Z., Angerer, R. C. & Angerer, L. M. 2006 A database of mRNA expression patterns for the sea urchin embryo. *Dev. Biol.* **300**, 476–484. (doi:10.1016/j.ydbio.2006.08.034)

31 Harris, D. J. & Kolen, M. J. 1988 Bootstrap and traditional standard errors of the point-biserial. *Educ. Psychol. Meas.* **48**, 43–51. (doi:10.1177/0013164488048001006)

32 Glass, G. V. & Hopkins, K. D. 1996 *Statistical methods in education and psychology*. Newton, MA: Allyn & Bacon.

33 Ashburner, M. *et al.* 2000 Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)

34 Bairoch, A. *et al.* 2009 The universal protein resource (UniProt). *Nucleic Acids Res.* **37**, D169–D174. (doi:10.1093/nar/gkn664)

35 Schmidt, P. S. & Rand, D. M. 2001 Adaptive maintenance of genetic polymorphism in an intertidal barnacle: habitat-and life-stage-specific survivorship of MPI genotypes. *Evolution* **55**, 1336–1344.

36 Marshall, D., Monro, K., Bode, M., Keough, M. & Swearer, S. 2010 Phenotype-environment mismatches reduce connectivity in the sea. *Ecol. Lett.* **13**, 128–140. (doi:10.1111/j.1461-0248.2009.01408.x)

37 Storey, J. & Tibshirani, R. 2003 Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445. (doi:10.1073/pnas.1530509100)

38 Lester, S., Tobin, E. & Behrens, M. 2007 Disease dynamics and the potential role of thermal stress in the sea urchin, *Strongylocentrotus purpuratus*. *Can. J. Fish. Aquat. Sci.* **64**, 314–323. (doi:10.1139/f07-010)

39 Britten, R. J., Cetta, A. & Davidson, E. H. 1978 The single-copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. *Cell* **15**, 1175–1186. (doi:10.1016/0092-8674(78)90044-2)

40 Tschopp, J., Martinon, F. & Burns, K. 2003 NALPs: a novel protein family involved in inflammation. *Nat. Rev. Mol. Cell Biol.* **4**, 95–104. (doi:10.1038/nrm1019)

41 Sanford, E. 2010 Local adaptation in the sea. *Annu. Rev. Mar. Sci.* **3**, 509–535.

42 Ciechanover, A., Orian, A. & Schwartz, A. 2000 Ubiquitin-mediated proteolysis: biological regulation via destruction. *Bioessays* **22**, 442–451. (doi:10.1002/(SICI)1521-1878(200005)22:5<442::AID-BIES6>3.0.CO;2-Q)

43 Thomas, J. 2006 Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res.* **16**, 1017–1030. (doi:10.1101/gr.5089806)

44 Townsend, J., Cavalieri, D. & Hartl, D. 2003 Population genetic variation in genome-wide gene expression. *Mol. Biol. Evol.* **20**, 955–963. (doi:10.1093/molbev/msg106)

45 Pearse, J., Costa, D., Yellin, M. & Agegian, C. 1977 Localized mass mortality of red sea urchin, *Strongylocentrotus franciscanus* near Santa Cruz, California. *Fish. Bull.* **75**, 645–648.

46 Lafferty, K. 2004 Fishing for lobsters indirectly increases epidemics in sea urchins. *Ecol. Appl.* **14**, 1566–1573. (doi:10.1890/03-5088)

47 Hibino, T. *et al.* 2006 The immune gene repertoire encoded in the purple sea urchin genome. *Dev. Biol.* **300**, 349–365. (doi:10.1016/j.ydbio.2006.08.065)

48 Roach, J., Glusman, G., Rowen, L., Kaur, A., Purcell, M., Smith, K., Hood, L. & Aderem, A. 2005 The evolution of vertebrate Toll-like receptors. *Proc. Natl Acad. Sci. USA* **102**, 9577–9582. (doi:10.1073/pnas.0502272102)

49 Hughes, A. & Nei, M. 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170. (doi:10.1038/335167a0)

50 Ferrer-Admetlla, A. 2008 Balancing selection is the main force shaping the evolution of innate immunity genes. *J. Immunol.* **181**, 1315.

51 Kurtz, J., Kalbe, M., Aeschlimann, P., Hberli, M., Wegner, K., Reusch, T. & Milinski, M. 2004 Major histocompatibility complex diversity influences parasite resistance and innate immunity in sticklebacks. *Proc. R. Soc. Lond. B* **271**, 197–204. (doi:10.1098/rspb.2003.2567)

52 Terwilliger, D. P., Buckley, K. M., Mehta, D., Moorjani, P. G. & Smith, L. C. 2006 Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*. *Physiol. Genomics* **26**, 134–144. (doi:10.1152/physiolgenomics.00011.2006)

53 Harvell, C., Mitchell, C., Ward, J., Altizer, S., Dobson, A., Ostfeld, R. & Samuel, M. 2002 Climate warming and disease risks for terrestrial and marine biota. *Science* **296**, 2158–2162. (doi:10.1126/science.1063699)

54 Britten, R. J. & Davidson, E. H. 1971 Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46**, 111–138.

55 King, M. & Wilson, A. 1975 Evolution at two levels in humans and chimpanze. *Science* **188**, 107–116. (doi:10.1126/science.1090005)

56 Wray, G. 2007 The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216. (doi:10.1038/nrg2063)

57 Carroll, S. B. 2008 Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36. (doi:10.1016/j.cell.2008.06.030)