Supplemental Text

We found enrichment for high $F_{st}$ values in a number of regions immediately upstream (2000bp) of annotated translation start sites, most notably in regions upstream of E3 ligases (tables 1 and 2). These signals may be driven by linkage between upstream variants and variants located in flanking, coding regions of the genome. However, as LD blocks are generally short in sea urchins, regulatory variants may be associated with potentially functional variants within more immediately flanking, non-coding sequence.

The distribution of regulatory elements
A number of studies have looked empirically and computationally for the distribution of functional regulatory regions across the genome in relationship to annotated coding genes. Though regulatory regions can be found across the genome, there is a significant enrichment for transcription factor binding sites in the first few KB immediately upstream of translation start sites and in 5'UTRs in humans (1-3) and other model systems (4, 5). These analyses, based largely on observed or predicted transcription factor binding have been further supported by studies making use of enhancer traps and DNAase hypersensitivity assays (6, 7). Experimental research thus supports the assertion that while distal regulatory enhancers can and do play important roles in regulating gene expression (8), upstream sequence is clearly enriched for regulatory elements. These results suggest, though do not prove, that variants underlying our upstream probes may cover, or be linked to, functional cis-regulatory variation.

In sea urchins, less is known about genome-wide distributions of regulatory elements. However, extensive information is available about the necessary and sufficient cis-regulatory regions underlying the expression of dozens of genes employed during sea urchin development (for references see http://sugp.caltech.edu/endomes/#References). In most of these cases, the majority of regulatory sites are found within a few kb upstream of the translation start site (9).

Our probes for both coding and non-coding regions of the genome were designed using the *S. purpuratus* genome sequence (10). However, many analyses of sea urchin promoters were carried out before the publication of the genome. For eight of these promoters (those underlying the expression of CyIIa, CyIIIa, Endo16, FoxB, HE, sm30, sm50, and Sp-AN (11-18)) we have aligned promoter sequences shown experimentally to recapitulate proper embryonic expression in reporter assays with 3kb immediately
upstream of the translation start site taken from the *S. purpuratus* v2.5 assembly (electronic supplemental material, dataset S2). In all cases, annotated regulatory regions are either directly covered by the regions to which we designed probes (2000 bp) or close enough to them to be in likely linkage disequilibrium.

**An enrichment of promoter-specific sequence motifs in upstream regions covered by array probes**

The identification of functional regulatory sites within non-coding DNA remains a significant computational challenge. However, because transcription factor binding sites often occur multiple times within a regulatory element, because transcription factor binding sites are often clustered, and because co-regulated and functionally similar genes are often regulated by similar suites of transcription factors, regulatory regions often show enrichments of sequence motifs that can be revealed by algorithmic analyses (19).

We made use of the popular motif discovery algorithm MEME (20) to look for similarities in enriched motifs in three sets of data: 1) 2000bp upstream of the E3-ligases showing significant elevation of *F*<sub>ST</sub> taken from the v2.5 assembly. 2) 2000bp upstream of the translation start of the 8 well-studied genes above also taken from the v2.5 assembly. 3) The validated *cis*-regulatory promoters for the eight well-studied gene as taken from the pre-genome assembly literature. For each of these datasets, we ran MEME with a minimum motif width of 6 and a maximum of 14. Background nucleotide frequencies (for p-value calculations) were obtained using a 1st order Markov model of the background nucleotide frequencies for each region. The same analyses were also conducted using 3000bp upstream rather than 2000bp to nearly identical results (data not shown).

The three most significantly enriched motifs are given in figure S2a,c,e. Similar motifs are identified in all three datasets, most notably a strong enrichment for a 14bp G-rich tract (a-motif1, c-motif2, e-motif1). Importantly, this motif is not seen when MEME was run on shuffled promoter sequences (electronic supplemental material, figure S2b,d,f) demonstrating that the presence of the motif is independent of base-pair composition *per se*.

The program STAMP (21) shows significant similarities between all three of the G-rich motifs and the recognition sites to RREB1, Pax4, SPQ, Klf4, and RPN4 (p<1e-05). However, this motif may not be directly related to transcription factor binding. A recent analysis of putative regulatory regions in humans and mice (22) showed that the most significantly enriched motif within 200bp of the annotated transcription start site in humans and mice was, interestingly, AGGGGGCGGGG. The biological meaning of this sequence is unclear, but that nearly the same motif was found in our data sets and an independent analysis of putative human and mouse regulatory regions (but not in the randomized data) suggests that our upstream probes overlapped functional regulatory sites.
Supplemental References


### Table S1. Summary of the number of categories tested, the number significant at q < 0.01, the critical p-value, and the potential number of false positives.

<table>
<thead>
<tr>
<th></th>
<th>F&lt;sub&gt;ST&lt;/sub&gt;</th>
<th></th>
<th>H&lt;sub&gt;T&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coding</td>
<td>upstream</td>
<td>coding</td>
</tr>
<tr>
<td></td>
<td>all adult larvae</td>
<td>all adult larvae</td>
<td>all adult larvae</td>
</tr>
<tr>
<td>No. categories tested</td>
<td>1196 610 398</td>
<td>470 202 139</td>
<td>1196 610 398</td>
</tr>
<tr>
<td>No. significant, q&lt;0.01</td>
<td>0 0 0</td>
<td>2 6 10</td>
<td>48 0 0</td>
</tr>
<tr>
<td>Critical p-value</td>
<td>- - -</td>
<td>2.92E-06 0.0003 0.0002</td>
<td>0.0004 - -</td>
</tr>
<tr>
<td>No. false positives</td>
<td>- - -</td>
<td>0.0014 0.0568 0.0276</td>
<td>0.4303 - -</td>
</tr>
</tbody>
</table>
Figure S1. Confirmation of allele frequency data with sequence data sampling an additional 10 individuals per population at six randomly selected polymorphic RSTA sites.
Figure S2. Sequence logos for the top three most highly enriched sequence motifs from three sets of sequences (and their randomized data). a,(b). 2000bp upstream of the E3-ligases showing significant elevation of $F_{ST}$ taken from the v2.5 assembly (table 2). c,(d). 2000bp upstream of the translation start of the 8 well-studied genes taken from the v2.5 assembly. e,(f). Validated cis-regulatory promoters for the eight well-studied gene as taken from the pre-genome assembly literature. Note the presence of a similar G-rich motif found in each of the datasets (a-Motif1, c-Motif1, e-Motif2). As the motif is not found in any of the randomized datasets (b,d,f), the presence of this sequence is independent of sequence composition per se.