Evidence That Purifying Selection Acts on Promoter Sequences

Robert K. Arthur and Ilya Ruvinsky[1]
Department of Ecology and Evolution, Committee on Genetics, Genomics and Systems Biology, Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637

ABSTRACT We tested whether functionally important sites in bacterial, yeast, and animal promoters are more conserved than their neighbors. We found that substitutions are predominantly seen in less important sites and that those that occurred tended to have less impact on gene expression than possible alternatives. These results suggest that purifying selection operates on promoter sequences.

THE study of *cis*-regulatory evolution presents "challenges beyond those typically encountered in analyses of coding sequence evolution" (Wray *et al.* 2003). We are currently unable to infer regulatory function from primary sequences and, consequently, do not have a clear understanding of a relationship between function and conservation. Whereas it is clear that many *cis*-elements are under selective constraint (Bergman and Kreitman 2001; Dermitzakis *et al.* 2003; Andolfatto 2005; Hahn 2007; Loots and Ovcharenko 2010), in some instances sites known to be functional in one species have been lost in closely related species (Ludwig *et al.* 1998; Dermitzakis and Clark 2002; Moses *et al.* 2006; Doniger and Fay 2007; Bradley *et al.* 2010). Genome annotation approaches, such as "phylogenetic footprinting" (Tagle *et al.* 1988; Blanchette and Tompa 2002; Zhang and Gerstein 2003) and "phylogenetic shadowing" (Boffelli *et al.* 2003), rely on greater conservation of functional sites compared to surrounding sequences, yet this supposition may not always be true (Emberly *et al.* 2003; Balhoff and Wray 2005). Indeed, positive selection may drive turnover of binding sites (Rockman *et al.* 2003; He *et al.* 2011). Although evidence suggests that fitness costs of mutations in noncoding regions may be relatively low (Kryukov *et al.* 2005; Chen *et al.* 2007; Raijman *et al.* 2008), few studies have explicitly tested the relationship between functions of individual nucleotides and the fitness costs of mutations at these sites (Shultzaberger *et al.* 2010).

Our knowledge of the forces driving the evolution of *cis*-elements largely comes from sequence comparisons between and within species, often without specific reference to the function of individual nucleotides within these elements (Wong and Nielsen 2004; Bush and Lahn 2006; Drake *et al.* 2006; Casillas *et al.* 2007). Yet regulatory functions and constraints are not uniformly distributed within *cis*-elements as evidenced by the correlation of conservation and functional importance of promoter motifs (Johnson *et al.* 2004).

Binding energy of transcription factor binding sites can be experimentally measured and computationally modeled (Djordjevic *et al.* 2003; Maerkl and Quake 2007; Weindl *et al.* 2007; Zhao *et al.* 2009). Modeling and comparative sequence analyses suggest that selection effects on binding sites may be mediated by their binding energy (Mustonen and Lässig 2005; Mustonen *et al.* 2008). Sites with high predicted binding strength appear to be more conserved, which is consistent with purifying selection (Moses 2009). Within transcription-factor-binding sites, substitutions occur at position-specific rates (Tanay *et al.* 2004; Kim *et al.* 2009). Specifically, the degree of conservation of individual nucleotides is proportional to their information content, likely because sites that make direct contact with transcription factors tend to be highly conserved (Mirny and Gelfand 2002; Moses *et al.* 2003). For this reason, it is tempting to use binding energy as a proxy for the functional consequences, and ultimately fitness effects, of mutations at a given site.

However, the relationship between binding, function, and fitness is not well understood (Mirny and Gelfand 2002). In

doi: 10.1534/genetics.111.133637
Manuscript received August 4, 2011; accepted for publication August 26, 2011
Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.111.133637/DC1.
[1]Corresponding author: University of Chicago, 1101 East 57th St., Chicago, IL 60637. E-mail: ruvinsky@uchicago.edu

some instances, there exists a correlation between binding energy and substitution rate (Brown and Callan 2004), but this may not always be the case (Kotelnikova *et al.* 2005). Furthermore, nonbinding nucleotides may exert some effect on transcription (Mai *et al.* 2000; Mirny and Gelfand 2002; Abnizova *et al.* 2007; Wozniak and Hughes 2008) and potentially on fitness.

A comprehensive understanding of the evolution of *cis*-regulatory sequences will require the synthesis of knowledge concerning binding energy, function, and fitness consequences of individual mutations within these elements. Because such data are not generally available, it would be desirable to ascertain whether a relationship exists between functions of specific nucleotides within *cis*-elements and their rates of evolution. Such analyses would constitute a critical link between functional studies and comparative sequence analysis.

Materials and Methods

Data

Functional data were derived from published articles reporting studies of promoter mutagenesis (Table S1). A "complete" data set for a given position would contain the information on the consequences of changing the wild-type nucleotide to every one of the three alternatives. Altogether, our data set contained 182 bp examined in such a way. There were also 209 nucleotides with "incomplete" data, *i.e.*, situations in which information was available for only one or two of the three possible substitutions. Although high-throughput mutagenesis data are available for several additional promoters (Patwardhan *et al.* 2009), we found that these data were inconsistent with the results of single-gene studies, even on the same promoter (data not shown). We therefore did not include them in the present analysis.

For each *cis*-regulatory element for which mutagenesis data were available, we identified orthologous sequences in a number of closely related species (Figure S1). In each broad taxonomic group (bacteria, yeast, animals), we endeavored to align sequence from a set of species of roughly equivalent phylogenetic distance (measured by the metric of substitutions per base pair). In counting substitutions, we took into account the phylogenetic relationship of the species being compared. For example, substitutions in sister species that could be parsimoniously attributed to the common ancestor of these species were counted once, not twice.

Statistical analyses

We calculated the "mutation cost index" for all experimentally characterized mutations; it is a measure of the extent to which a mutation alters promoter function. Expression levels of all mutagenized promoters were normalized to the expression levels of the wild-type promoter (in rare instances when the mutant promoter drove higher expression, the inverse of the normalization ratio was recorded instead). For a mutation that reduced gene expression to α (normalized to the level of the wild-type promoter), mutation cost index was defined as $1 - \alpha$; therefore, it can range from 0 (no alteration of expression level) to 1 (complete abrogation of promoter function). Similarly, every nucleotide within a promoter can be said to have a "site index," computed as a sum of the mutation cost indexes of all three possible substitutions. Site index can range from 0 (all mutations are inconsequential to promoter function) to 3 (all mutations at the site abolish expression). In the case of two promoters, a measure of function other than the level of gene expression was used (Table S1).

To test whether the mutation cost index was significantly lower for substitutions than for all possible mutations, as would be expected under purifying selection, we performed sampled randomization tests in which artificial data sets were generated by randomly sampling from the set of all mutations (Sokal and Rohlf 1995). Each artificially generated set matched the substitution data in the number of mutations, but differed in the specific mutations sampled. In the most general version of the test the artificially generated sets were randomly drawn from all experimentally tested mutations.

We performed three variations of this test, each of which constrained certain characteristics of the sampled sets. In the first, artificial sets were constructed to have the same frequency of nucleotides as that of the sites that sustained substitutions. In the second variation of the test, artificial sets were matched in nucleotide frequencies to those of derived nucleotides (*i.e.*, those nucleotides to which substitutions changed the ancestral nucleotides). In the third, the numbers of transition and transversion mutations were matched between the set of substitutions and the artificially generated sets. Sampled randomization tests were performed separately for bacteria, yeast, and animals. Each test was composed of at least 10,000 artificially generated sets. We calculated the fraction of instances in which the mean of an artificial set was lower than or equal to that of the substitutions set. This ratio, which represents the probability that the observed set of substitutions would occur by chance alone, constituted the reported *P*-value. We chose this method because the distributions of mutation cost indexes were highly non-normal and the substitutions represented a subset of all possible mutations. The sampled randomization test makes no assumptions about the underlying data. It reports the likelihood that the observed data set resembles a randomly chosen data set in regards to certain summary statistics (Sokal and Rohlf 1995). All statistical analyses were performed in the R statistical programming language (http://www.r-project.org).

Results

We explicitly tested whether functionally important nucleotides within promoters evolve under the same regime as their neighbors. A number of studies have been published in which individual nucleotides in a given promoter were

replaced while holding all other nucleotides constant (*e.g.*, Myers *et al.* 1985). Most commonly, mutagenized promoters were fused to reporter genes to compare their levels of expression to wild-type promoters. These tests measured the impact of each nucleotide substitution on the level of expression. For example, at a given site, an A could be a wild-type nucleotide, while mutations to C, G, and T could reduce gene expression to 10, 30, and 60% of the wild-type level, respectively. Combining these functional data with analysis of orthologous promoters could establish a relationship between function and rates of evolution. We assembled a data set of 14 such studies (Table S1), conducted on organisms from three distinct phylogenetic groups: animals (4), yeast (5), and bacteria (5). Together, these articles reported mutagenesis of 332 nucleotides and examined expression levels of 1040 constructs (animals: 136 nucleotides, 350 constructs; yeast: 79 nucleotides, 275 constructs; bacteria: 117 nucleotides, 415 constructs). Of all these experimentally tested nucleotides, 56 were inferred to have sustained substitutions (Figure S1). While limited in size, we believe that this data set is a near-exhaustive collection of published articles reporting experiments of this type.

It may be expected that the effects on gene expression of substitutions that accumulated during evolution would be less severe than the effects of average mutations that could have occurred within these promoters. We tested this hypothesis (Figure 1). We found that the mutations corresponding to substitutions had lower mutation cost indexes than average mutations (bacteria: $P = 1 \times 10^{-4}$; yeast: $P = 4 \times 10^{-4}$; animals: $P < 10^{-5}$). Therefore, among the substitutions that did occur, there was a substantial bias in favor of changes with lower impact on gene expression. This implies that purifying selection has acted to maintain gene expression levels.

Results in Figure 1 suggest that, in general, mutations with lower effects on promoter function tended to become fixed. Two distinct scenarios could account for this trend. First, the milder fixed substitutions could be distributed relatively evenly across sites. Alternatively, they may preferentially occur at a particular subset of sites. We used the functional data described above to test the hypothesis that substitutions are more common at sites where mutations have less severe effect on gene expression levels (Figure 2). One measure of functional importance of a site is an index defined as a sum of the mutation cost indexes of all three possible mutations that could occur at this nucleotide. Site indexes were significantly lower for positions with substitutions compared to all sites for which experimental mutagenesis data were available (bacteria: $P = 3.2 \times 10^{-3}$; yeast $= 6.4 \times 10^{-3}$; animals: $P = 3.4 \times 10^{-3}$). Therefore, in all three groups, substitutions preferentially occurred at sites that were less disruptive of gene expression.

Mutational biases are not sufficient to account for the trends reported above. First, in the promoter sequences that we analyzed there was no systematic difference in nucleotide composition between sites that sustained substitutions and those that did not (Table S2). Second, correcting for multiple hypothesis testing, there were no significant differences in mutation cost indexes between mutations involving different wild-type nucleotides (Figure S2). Finally, we repeated sampled randomization tests holding constant the number of (i) wild type and (ii) derived nucleotides and (iii) transitions and transversions. All of these modified tests showed significant differences between mutation cost indexes of substitutions compared to all possible mutations (Table S3).

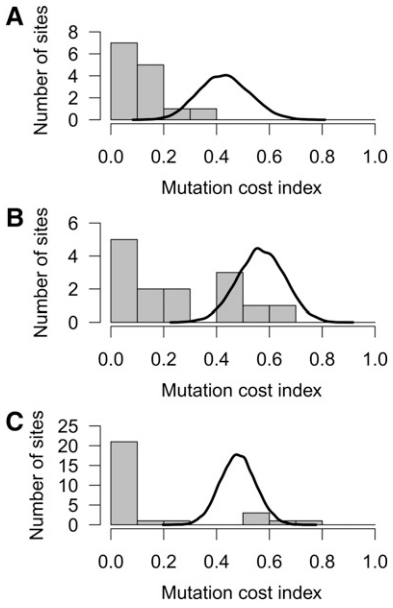

Figure 1 Substitutions in promoters have significantly milder effects on levels of gene expression than the mean effects of all possible mutations. Shaded bars depict mutation cost indexes of substitutions in (A) bacterial, (B) yeast, and (C) animal promoters. Line curves show the distribution of average mutation cost indexes obtained using a sampled randomization procedure.

Discussion

Our results suggest that purifying selection acts on promoter sequences in bacteria, yeast, and animals because we saw fewer than expected substitutions that corresponded to mutations of substantial effect. While these findings are concordant with previous reports of sequence conservation in *cis*-elements (Andolfatto 2005; Drake *et al.* 2006; Casillas *et al.* 2007; Molina and Van Nimwegen 2008), they add an important functional explanation for the observed patterns. An additional reason for the relative abundance of mutations of smaller effect is that they would be more likely to be beneficial and therefore be fixed by directional selection. Positive selection has been shown to act on *cis*-regulatory elements (Rockman *et al.* 2005; Haygood *et al.* 2007), and it may drive transcription-factor-binding-site turnover (Rockman *et al.* 2003; He *et al.* 2011). The inference of both positive and negative selection may not be contradictory, as it has been shown that both types of selection operate on gene regulatory

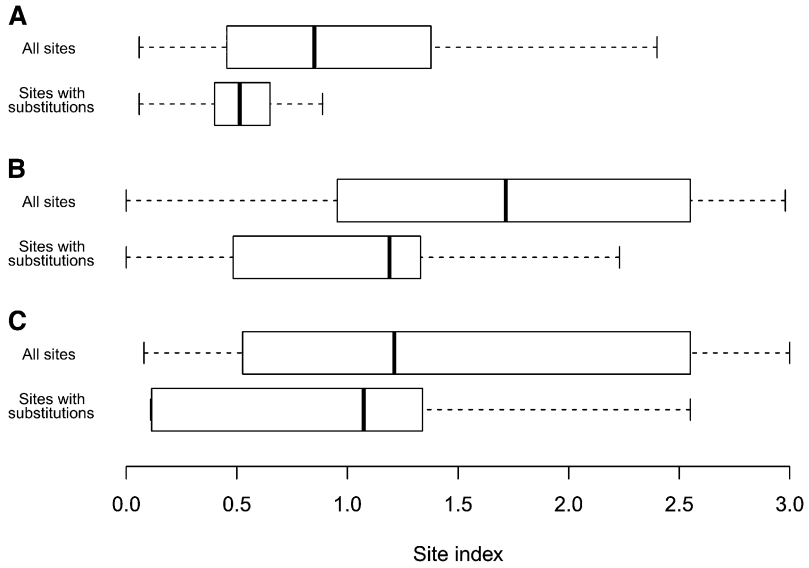

Figure 2 Substitutions in promoters tended to occur at sites with less severe impact on expression. Distributions of site indexes are shown for all sites and for sites with substitutions for (A) bacterial, (B) yeast, and (C) animal promoters.

elements in a variety of species (Kohn *et al.* 2004; Macdonald and Long 2005; Haddrill *et al.* 2008; Torgerson *et al.* 2009). Also, at least some regulatory regions are evolving under stabilizing selection (Ludwig *et al.* 2000; Loisel *et al.* 2006).

Five caveats should be noted. First, it is generally not known how changes in the level of expression translate into measures of fitness. However, our conclusions do not require a particular relationship, but merely a positive correlation between the extent to which a mutation changes expression of a gene and its fitness consequences. Available data suggest that such a correlation is likely (Shultzaberger *et al.* 2010). Second, the set of mutagenized sites was not random in all studies. In some cases, experimenters chose sites in which to induce mutations in a way presumably biased in favor of nucleotides expected to have more dramatic effects on gene expression. Third, the functional effects of mutations in promoters are highly context specific (Vidal *et al.* 1995). Therefore, fitness consequences of mutations are contingent on the backgrounds on which they occur and may have changed substantially over time (Bullaughey 2011). Fourth, functions of mutated promoters were tested either in cell lines (animals) or under laboratory conditions (bacteria and yeast). This leaves open a possibility that *in vivo* or under different environmental conditions, mutations seen in the laboratory as "functionally silent" may have substantial impact on fitness. Furthermore, although a point mutation of a given nucleotide may not have caused an appreciable change in expression level, the site may still be under selection because its deletion could cause a substantial decrease in gene expression (Patwardhan *et al.* 2009). It appears unlikely, however, that mutations that abrogate or substantially reduce expression are selectively neutral. Finally, all sequences analyzed in this study were derived from proximal promoter elements. The arrangement and composition of functional sites may be different between promoters and other *cis*-regulatory elements. Therefore, different types of *cis*-sequences may evolve under different selective regimes. Nonetheless, the results presented here highlight the value of functional data obtained at single-nucleotide resolution, not solely binding energy, for understanding regulatory evolution.

Acknowledgments

We are grateful to Kevin Bullaughey for numerous suggestions for improvement and help with data analysis. We thank Bin He and Marty Kreitman for critical reading and helpful suggestions and Chan Hee Choi for help during an early stage of this study. This work was made possible by grant support from the National Science Foundation (IOS-0843504) and the National Institutes of Health (NIH) (P50 GM081892) to I.R. and by an NIH training grant (T32 GM007197) to R.K.A.

Literature cited

Abnizova, I., T. Subhankulova, and W. Gilks, 2007 Recent computational approaches to understand gene regulation: mining gene regulation *in silico*. Curr. Genomics 8: 79–91.

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149–1152.

Balhoff, J. P., and G. A. Wray, 2005 Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc. Natl. Acad. Sci. USA 102: 8591–8596.

Bergman, C. M., and M. Kreitman, 2001 Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res. 11: 1335–1345.

Blanchette, M., and M. Tompa, 2002 Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res. 12: 739–748.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko *et al.*, 2003 Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391–1394.

Bradley, R. K., X. Y. Li, C. Trapnell, S. Davidson, L. Pachter *et al.*, 2010 Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol. 8: e1000343.

Brown, C. T., and C. C. Callan, 2004 Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 101: 2404–2409.

Bullaughey, K., 2011 Changes in selective effects over time facilitate turnover of enhancer sequences. Genetics 187: 567–582.

Bush, E. C., and B. T. Lahn, 2006 The evolution of word composition in metazoan promoter sequence. PLoS Comput. Biol. 2: e150.

Casillas, S., A. Barbadilla, and C. M. Bergman, 2007 Purifying selection maintains highly conserved noncoding sequences in Drosophila. Mol. Biol. Evol. 24: 2222–2234.

Chen, C. T., J. C. Wang, and B. A. Cohen, 2007 The strength of selection on ultraconserved elements in the human genome. Am. J. Hum. Genet. 80: 692–704.

Dermitzakis, E. T., and A. G. Clark, 2002 Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. 19: 1114–1121.

Dermitzakis, E. T., C. M. Bergman, and A. G. Clark, 2003 Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol. Biol. Evol. 20: 703–714.

Djordjevic, M., A. M. Sengupta, and B. I. Shraiman, 2003 A biophysical approach to transcription factor binding site discovery. Genome Res. 13: 2381–2390.

Doniger, S. W., and J. C. Fay, 2007 Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol. 3: e99.

Drake, J. A., C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh *et al.*, 2006 Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. 38: 223–227.

Emberly, E., N. Rajewsky, and E. D. Siggia, 2003 Conservation of regulatory elements between two species of Drosophila. BMC Bioinformatics 4: 57.

Haddrill, P. R., D. Bachtrog, and P. Andolfatto, 2008 Positive and negative selection on noncoding DNA in *Drosophila simulans*. Mol. Biol. Evol. 25: 1825–1834.

Hahn, M. W., 2007 Detecting natural selection on *cis*-regulatory DNA. Genetica 129: 7–18.

Haygood, R., O. Fedrigo, B. Hanson, K. D. Yokoyama, and G. A. Wray, 2007 Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat. Genet. 39: 1140–1144.

He, B. Z., A. K. Holloway, S. J. Maerkl, and M. Kreitman, 2011 Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet. 7: e1002053.

Johnson, D. S., B. Davidson, C. D. Brown, W. C. Smith, and A. Sidow, 2004 Noncoding regulatory sequences of *Ciona* exhibit strong correspondence between evolutionary constraint and functional importance. Genome Res. 14: 2448–2456.

Kim, J., X. He, and S. Sinha, 2009 Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet. 5: e1000330.

Kohn, M. H., S. Fang, and C. I. Wu, 2004 Inference of positive and negative selection on the 5′ regulatory regions of Drosophila genes. Mol. Biol. Evol. 21: 374–383.

Kotelnikova, E. A., V. J. Makeev, and M. S. Gelfand, 2005 Evolution of transcription factor DNA binding sites. Gene 347: 255–263.

Kryukov, G. V., S. Schmidt, and S. Sunyaev, 2005 Small fitness effect of mutations in highly conserved non-coding regions. Hum. Mol. Genet. 14: 2221–2229.

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

Loots, G. G., and I. Ovcharenko, 2010 Human variation in short regions predisposed to deep evolutionary conservation. Mol. Biol. Evol. 27: 1279–1288.

Ludwig, M. Z., N. H. Patel, and M. Kreitman, 1998 Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. Development 125: 949–958.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.

Macdonald, S. J., and A. D. Long, 2005 Identifying signatures of selection at the *enhancer of split* neurogenic gene complex in *Drosophila*. Mol. Biol. Evol. 22: 607–619.

Maerkl, S. J., and S. R. Quake, 2007 A systems approach to measuring the binding energy landscapes of transcription factors. Science 315: 233–237.

Mai, X., S. Chou, and K. Struhl, 2000 Preferential accessibility of the yeast his3 promoter is determined by a general property of the DNA sequence, not by specific elements. Mol. Cell. Biol. 20: 6668–6676.

Mirny, L. A., and M. S. Gelfand, 2002 Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res. 30: 1704–1711.

Molina, N., and E. van Nimwegen, 2008 Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res. 18: 148–160.

Moses, A. M., 2009 Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol. Biol. 9: 286.

Moses, A. M., D. Y. Chiang, M. Kellis, E. S. Lander, and M. B. Eisen, 2003 Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol. Biol. 3: 19.

Moses, A. M., D. A. Pollard, D. A. Nix, V. N. Iyer, X.-Y. Li *et al.*, 2006 Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput. Biol. 2: e130.

Mustonen, V., and M. Lässig, 2005 Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc. Natl. Acad. Sci. USA 102: 15936–15941.

Mustonen, V., J. Kinney, C. C. Callan, and M. Lassig, 2008 Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proc. Natl. Acad. Sci. USA 105: 12376–12381.

Myers, R. M., L. S. Lerman, and T. Maniatis, 1985 A general method for saturation mutagenesis of cloned DNA fragments. Science 229: 242–247.

Patwardhan, R. P., C. Lee, O. Litvin, D. L. Young, D. Pe'er *et al.*, 2009 High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. 27: 1173–1175.

Raijman, D., R. Shamir, and A. Tanay, 2008 Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. PLoS Comput. Biol. 4: e7.

Rockman, M. V., M. W. Hahn, N. Soranzo, D. B. Goldstein, and G. A. Wray, 2003 Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. 13: 2118–2123.

Rockman, M. V., M. W. Hahn, N. Soranzo, F. Zimprich, D. B. Goldstein *et al.*, 2005 Ancient and recent positive selection transformed opioid *cis*-regulation in humans. PLoS Biol. 3: e387.

Shultzaberger, R. K., D. S. Malashock, J. F. Kirsch, and M. B. Eisen, 2010 The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. PLoS Genet. 6: e1001042.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*. W. H. Freeman, New York.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess *et al.*, 1988 Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. 203: 439–455.

Tanay, A., I. Gat-Viks, and R. Shamir, 2004 A global view of the selection forces in the evolution of yeast *cis*-regulation. Genome Res. 14: 829–834.

Torgerson, D. G., A. R. Boyko, R. D. Hernandez, A. Indap, X. Hu *et al.*, 2009 Evolutionary processes acting on candidate *cis*-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet. 5: e1000592.

Vidal, M., A. M. Buckley, C. Yohn, D. J. Hoeppner, and R. F. Gaber, 1995 Identification of essential nucleotides in an upstream repressing sequence of *Saccharomyces cerevisiae* by selection for increased expression of TRK2. Proc. Natl. Acad. Sci. USA 92: 2370–2374.

Weindl, J., P. Hanus, Z. Dawy, J. Zech, J. Hagenauer *et al.*, 2007 Modeling DNA-binding of *Escherichia coli sigma70* exhibits a characteristic energy landscape around strong promoters. Nucleic Acids Res. 35: 7003–7010.

Wong, W. S., and R. Nielsen, 2004 Detecting selection in noncoding regions of nucleotide sequences. Genetics 167: 949–958.

Wozniak, C. E., and K. T. Hughes, 2008 Genetic dissection of the consensus sequence for the class 2 and class 3 flagellar promoters. J. Mol. Biol. 379: 936–952.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 20: 1377–1419.

Zhang, Z., and M. Gerstein, 2003 Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J. Biol. 2: 11.

Zhao, Y., D. Granas, and G. D. Stormo, 2009 Inferring binding energies from selected binding sites. PLoS Comput. Biol. 5: e1000590.

Communicating editor: D. Begun

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.111.133637/DC1

Evidence That Purifying Selection Acts on Promoter Sequences

Robert K. Arthur and Ilya Ruvinsky

DOI: 10.1534/genetics.111.133637

A. Bacteria

Sigma-x promoter: 3 substitutions; 26 nucleotides

Bacillus subtilis	TGTAATGTAACTTTTCAAGCTATTCATACGACAA
Bacillus sp. BT1B_CT2	TGTAATGTAACCTTTTTAAGATTGACAAACGACAA
Bacillus subtilis subsp. spizizenii ATCC 6633	TGTAATGTAACTTTTCAAGCTATTAATACGACAA

Chlamydia trachomatis rRNA promoter : 6 substitutions; 41 nucleotides

Chlamydia trachomatis 16s rRNA	AAAAATAGATGCAGAAAAAATAGAGGTTGATATAAGATGTT
C.trach L2tet1WGS	AAAAAAAGGTGCAAAAAAAAATAGGGGGTGACATAAGATGTT
CmuridarumMopnTet14WGS	AAAAAAAGGTGCAAAAAAAAATAGGGGGTGACATAAGATGTT

Escherischia coli carAB: 1 substitution; 8 nucleotides

Escherischia coli	ATATTCTCT
Candidatus Blochmannia floridanus	ATATTGTGT
Salmonella enterica subsp. Enterica serovar Typhimurium str. LT2	ATATTCTCT
Shigella boydii Sb227	ATATTCTCT

Escherischia coli rRNA: 1 substitution; 25 nucleotides

Escherischia coli	TTTTAAATTTCCTCTTGTCAGGCCGGAATAACTCCCTATAATGCGCCACCAC
Salmonella enterica subsp. Enterica serovar Tennessee str. CDC07-0191	TTTTAAATTTCCTCTTGTCAGGCAGAAATAACTCCCTATAATGCGCCACCAC
Enterobacteriaceae bacterium 9_2_54FAA	TTTCAAATAAACACTTGTCAGCCGTTCAGAAGTCCCTATAATGCGCCACCAC

Salmonella typhimurium strain LT2: 3 substitutions; 17 nucleotides

Salmonella typhimurium	TCAAGTCC	TGCCGATAA
Enterobacter sp. 638	TCAAGTTT	TGTCGATAA

B. Yeast

ARG3: 1 substitution; 22 nucleotides

Saccharomyces cerevisiae	CTTTAAGTACAGTTAATAACGAGC
Saccharomyces paradoxus	CTTTAAGTACAGTTAATAACGAGC
Saccharomyces mikatae	CTTTAAGTACAGTTAATAACGAGC
Saccharomyces kudriavzevii	CTCTAAGTACAGTTAATAACGAGC
Saccharomyces bayanus	CTTTAAGTACAGTTGATAACGAGC

CAR1: 2 substitutions; 13 nucleotides

Saccharomyces cerevisiae	GTAGCCGCCGAGG
Saccharomyces mikatae	GTAGCCGCCGAGA
Saccharomyces kudriavzevii	GTAGCCGCCGAAA
Saccharomyces bayanus	GTAGCCGCCGAGA
Saccharomyces paradoxus	GTAGCCGCCGAGG

DAL5: 6 substitutions; 15 nucleotides

Saccharomyces cerevisiae	TTGCTGATAAGGTGC
Saccharomyces kudriavzevii	TTGCTGATAAGGTAC
Saccharomyces bayanus	TTGCTGATAAGATGC
Saccharomyces mikatae	TTGCTGATAAGCAAC
Saccharomyces paradoxus	TTACTGATAAGGTGC

His3: 2 substitutions; 6 nucleotides

Saccharomyces cerevisiae	TATAAA
Saccharomyces cariocanus	TATAAG
Saccharomyces kudriavzevii	TATAAA
Saccharomyces mikatae	TATAAG
Saccharomyces bayanus	TATAAA

STE6: 3 substitutions; 23 nucleotides

Saccharomyces cerevisiae	CATGTAATTACCTAATAGGGAAATTTACAC
Saccharomyces kudriavzevii	CATGTAATTACCTAATTAGGAAATTTACAC
Saccharomyces paradoxus	CATGTAATTACCTAATTCGGAAATTTACAC
Saccharomyces bayanus	CATGTAATTACCTAATCCGGAAATTTACAC
Saccharomyces mikatae	CATGTAGTTACCAAATTAGGAAATTTACAC

C. Animals

Vav: 1 substitution; 21 nucleotides

Homo sapiens	CAGGCAAAGAAGAGGAAGTGG
Canis lupus	CAGGCAAAGAAGAGGAAGTGG
Felis catus	CAGGCAAAGAGGAAGTGG
Macaca mulatta	CAGGCAAAGAAGAGGAAGTGG
Rattus norvegicus	CAGTCATAGAAGAGGAAGTGG
Mus musculus	CAGTCACAGAAGAGGAAGTGG

HSP70.1: 3 substitutions; 14 nucleotides

Homo sapiens	GGAATATTCCCGAC
Macaca mulatta	GGAATATTCCCGAC
Rattus norvegicus	GGAAGATTCCTGGC
Canis lupus	GGAATCTTCCCGAC
Felis catus	GGAATATTCCCGGC
Mus musculus	GGAAGATTCCTGGC

B-globin: 21 substitutions; 92 nucleotides

Mus musculus

AGAGCCACACCCTGGTAAGGGCCAATCTGCTCACACAGGATAGAGAGGGCAGGAGCCAGGGCAGAGCATATAAGGTGAGGTAGGATCAGT

TGCTCCTCACATTTGCTTCTGACA

Rattus norvegicus

AGAGCCACACCCTGGTATTGGCCAATCTGCTCACACAGGACAGCGAGAGCAGGAGCCAGGCAGAGCATAAAAGGTGGGGCGGATCAGT

CGCTCCTCACATTTGCTTCTGACA

Homo sapiens

GGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCAT

CTATTGCTTACATTTGCTTCTGACA

Felis catus

CCACACCTAGGCCTGGGCCAATCTGCTCACAGGAGCAGGGAGGGTAAGATCAGGCCT

GGGCATAAAAGGAAGAGCAGGGATAGCTACCAGCTTACACTTGCTTCTGACA

Macaca mulatta

GGAGCCACACCCTACAGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCAT

CTATTGCTTACACTTGCTTCTGACA

<u>Drosophila tRNA: 3 substitutions; 9 nucleotides</u>

Drosophila melanogaster	GGTTCGAGTCC
Drosophila simulans	GGTTCGAGTCC
Drosophila sechellia	GGTTCGAGTCC
Drosophila anassae	GGTAGGAGTCC
Drosophila willistoni	GGTTGGAGTCC
Drosophila erecta	GGTTCCAGTCC

Figure S1 Alignments of mutagenized regions of promoters analyzed in this study. Alignments are separated by phylogenetic group and gene. The species in which the mutagenesis experiment was carried out is listed first. The number of nucleotides refers to a total number of sites mutagenized in the experimental study. A subset of these positions have sustained substitutions; these are boxed. In some cases, the number of boxed nucleotides is greater than the number of substitutions listed in the title line. In such instances, information on the functional consequences of the particular mutation that corresponded to a substitution was not available in the experimental study.

Figure S2 Mutation cost indexes are indistinguishable for various substitutions. Wild type nucleotides are shown in large letters on the left and derived nucleotides are shown as small letters beside. Every possible comparison of the wild type distributions against each other was performed using the Kolmogorov-Smirnov test. The results, as p-values, are: A-C: 0.01; A-G: 0.24; A-T: 0.20; C-G: 0.16; C-T: 0.03; G-T: 0.10. Note that the Kolmogorov-Smirnov test produces approximate p-values in the case of tied values, which were present, and that when a Bonferroni correction for multiple-testing was applied, none of the above p-values were significant.

Table S1 A list of published articles reporting promoter mutagenesis experiments that were analyzed in this study

Group	Organism	Gene	Citation	Assay
Yeast	S. cerevisiae	ARG3	De Rijcke *et al.* (1992) *Mol. Cell. Biol.* 12(1): 68-81.	OTCase
Yeast	S. cerevisiae	CAR1	Luche *et al.* (1990) *Mol. Cell. Biol.* 10(8): 3884-95.	Beta-galactosidase
Yeast	S. cerevisiae	his3	Chen *et al.* (1988) *Proc. Natl. Acad. Sci. USA* 85(8): 2691-5.	Beta-galactosidase
Yeast	S. cerevisiae	Suc2	Lundin *et al.* (1994) *Mol. Cell. Biol.* 14(3): 1979-85.	Beta-galactosidase
Yeast	S. cerevisiae	Dal5	Bysani *et al.* (1991) *J. Bact*. 173(16): 4977-82.	Beta-galactosidase
Bacteria	E. coli	carAB	Wang *et al.* (1998) *J. Mol. Biol.* 277(4): 805-24.	Occupancy
Bacteria	B. subtilis	Abh, RapD, LytR	Huang *et al.* (1998) *J. Mol. Biol.* 279(1): 165-73.	Beta-galactosidase
Bacteria	C. trachomatis	rRNA	Tan *et al.* (1998) *J. Bact*. 180(9): 2359-66.	In vitro transcription
Bacteria	E. coli	rRNA	Gaal *et al.* (1989) *J. Bact*. 171(9): 4852-61.	Beta-galactosidase
Bacteria	S. typhimurium	flgKL	Wozniak *et al.* (2008) *J. Mol. Biol.* 379(5): 936-952.	Beta-galactosidase
Animals	H. sapiens	HSP70.1	Cunniff *et al.* (1993) *J. Biol. Chem*. 268(11): 8317-24.	Primer extension
Animals	H. sapiens	vav	Denkinger *et al.* (2002) *Reactions* 783: 772-783.	Competitive EMSA
Animals	M. musculus	B-globin	Myers *et al.* (1986) *Science* 232: 613-618.	Beta-galactosidase
Animals	D. melanogaster	tRNA	Gaëta *et al.* (1990) *Nucl. Acids Res.* 18(6): 1541-8.	In vitro transcription

Table S2 Expected and observed nucleotide compositions in promoters analyzed in this study

Bacteria	Expected	Observed
A	5.03	2
C	2.75	6
G	2.10	1
T	4.12	5
p-value	**0.5161**	

Yeast	Expected	Observed
A	5.18	6
C	2.24	0
G	3.36	7
T	3.22	1
p-value	**0.2295**	

Animals	Expected	Observed
A	8.32	7
C	6.49	6
G	9.13	10
T	4.06	5
p-value	**1**	

The number of expected and observed nucleotides, separated by group. The 'expected' number was calculated by multiplying the proportion of each base in the full data set by the number of substitutions (for each group). The p-value, obtained by comparing expected and observed columns with Fisher's Exact Test, is shown in bold beneath each group's data.

Table S3 Results of sampled randomization tests performed to account for a number of mutational biases

Test for mutant cost indexes	Bacteria	Yeast	Animals
no stratification	1×10^{-4}	1×10^{-4}	$< 1 \times 10^{-5}$
by wild type nucleotide	1×10^{-4}	4×10^{-4}	$< 1 \times 10^{-5}$
by derived nucleotide	2.0×10^{-4}	1.9×10^{-3}	$< 1 \times 10^{-5}$
by transition/transversion	1.9×10^{-3}	1.3×10^{-3}	$< 1 \times 10^{-5}$
Test for site indexes			
no stratification	1.4×10^{-2}	2.1×10^{-2}	4.4×10^{-2}
by wildtype nucleotide	3.2×10^{-3}	6.4×10^{-3}	3.4×10^{-3}

The results of each sampled randomization test when controlling for various factors. For each test, the substitution set is made to agree with the samples obtained via randomized sampling with respect to one factor. For instance, controlling for the number of wild type nucleotides means that the number of substitutions at sites which are originally A, T, G, C are retained in the samples, such that each test result draws a corresponding set of mutations which started as those nucleotides. No stratification means that all factors were ignored, and a completely random set of mutations was drawn.