Cis-regulatory Elements and Human Evolution

Adam Siepel*, Leonardo Arbiza*

Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

Abstract

Modification of gene regulation has long been considered an important force in human evolution, particularly through changes to *cis*-regulatory elements (CREs) that function in transcriptional regulation. For decades, however, the study of *cis*-regulatory evolution was severely limited by the available data. New data sets describing the locations of CREs and genetic variation within and between species have now made it possible to study CRE evolution much more directly on a genome-wide scale. Here, we review recent research on the evolution of CREs in humans based on large-scale genomic data sets. We consider inferences based on primate divergence, human polymorphism, and combinations of divergence and polymorphism. We then consider "new frontiers" in this field stemming from recent research on transcriptional regulation.

Keywords: transcriptional regulation, divergence, polymorphism, population genomics

Introduction

The chimpanzee has long presented a conundrum for human geneticists. The orthologous proteins of humans and chimpanzees are more than 99.5% identical [1], yet the two species differ profoundly across a broad spectrum of apparently unrelated phenotypes. This evident paradox led King and Wilson to speculate, famously, that differences in gene regulation, rather than protein-coding sequences, might primarily explain differences in physiology and behavior between humans and chimpanzees [2] (see also [3, 4]). This proposal—while bold—in a sense grew naturally out of Jacob and Monod's research over a decade earlier establishing that the "program" for gene regulation was, in large part, written in DNA [5]. For, as Jacob and Monod themselves recognized [6], if regulatory programs were encoded in the genome, then they were subject to modification by mutation and natural selection, just as protein structure was.

These early conjectures about regulatory evolution were alluring, but for a long time they remained frustratingly abstract and unsubstantiated. In those days, few details could be provided about precisely which regulatory sequences changed, how much, and with what effect. During the ensuing decades, however, indirect evidence and anecdotal examples began to accumulate in support of the idea that *cis*-regulatory elements (CREs) associated with transcriptional regulation played a particularly central role in regulatory evolution [7–9]. (For the purposes of this article, CREs are regulatory sequences relatively near their target gene, typically no more than about a megabase from the transcription unit; we will focus on CREs involved in transcription.) Nevertheless, direct, large-scale support for the prominence of CREs in the evolution of form and function was lacking, and these claims remained controversial [10].

During the past few years, it has finally become possible to examine the evolution of CREs directly on a genome-wide scale, owing to the availability of genomic data describing both genetic variation and

*Corresponding author: Siepel, Adam (acs4@cornell.edu) regulatory elements. This review will cover major developments over the past decade in the study of human CREs and their role in human evolution, with a particular focus on studies that have leveraged the large public data sets released over the past 2–3 years. Along the way, we will discuss various challenges that arise in the interpretation of these data sets. We will end with a brief survey of new developments in the study of transcriptional regulation that have the potential to enrich studies of human evolution.

The Old Wave: Studies Based on Interspecies Divergence

A central principle of molecular evolution holds that inferences about natural selection can be made by comparing rates of nucleotide substitution in sites of functional importance with those at sites expected to have little or no influence on fitness. This principle is based on the expectation that mutations will occur at approximately equal rates in both functional and nonfunctional sites, but natural selection will alter the rates at which derived alleles reach fixation in functional sites (Figure 1). This idea has been applied for decades to protein-coding sequences, where amino acid altering (nonsynonymous) and non-altering (synonymous) substitutions provide convenenient classes to contrast [11–13].

The sequencing of the chimpanzee genome [1] enabled analogous methods to be applied genome-wide to putative CREs in hominids. For example, Keightley et al. examined sequences in upstream regions and first introns of genes and contrasted them with other intronic sequences assumed to be neutrally evolving [14]. They found that putative regulatory sequences showed almost no evidence of constraint in hominids, but were significantly constrained in mouse and rat. Finding no signs of positive selection, they argued that regulatory sequences in hominids had experienced "widespread degradation" due to their reduced effective population sizes (see also [1, 15, 16]). Soon afterward, Khaitovich et al. analyzed human-chimpanzee divergence patterns in promoter regions together with data on mRNA expression. Interestingly, they found that human-chimpanzee divergence in gene expression (normalized for intraspecies diversity) was much more pronounced in the testis than in the brain or several other tissues, possibly reflecting positive selection due to differences in mating strategies. They did find an excess of lineage-specific changes in expression of brain genes in human relative to chimpanzee.

Haygood and colleagues improved on the statistical methology of previous studies by developing a phylogenetic likelihood ratio test analogous to those used for protein-coding sequences [17, 18] for lineage-specific elevations in substitution rates in promoter regions [19] (see Figure 1C). Based on alignments of the human, chimpanzee, and rhesus macaque genome sequences, Haygood et al. found evidence of positive selection acting on the promoters of at least 250 genes. High-scoring genes were significantly enriched for roles in neural development and function, nutrition, and metabolism, suggesting an important role for CREs in human cognitive, behavioral, and dietary adaptations. Another series of studies, based on similar statistical methods, tested conserved noncoding sequences for "accelerated" evolution in humans [20–23].

The first large-scale study of primate evolution to make use of newly emerging chromatin immunoprecipitation and microarray (ChIP-chip) data for TF binding was carried out by Gaffney and colleagues [24]. The authors collected ChIP-chip data from seven previously published studies, and then analyzed patterns of divergence at bound sites in the human, chimpanzee, and rhesus macaque genomes, comparing the regulatory sequences with "control" regions. They also considered transcription factor binding sites (TFBSs) recorded in the TRANSFAC database. Using a simple divergence-based estimator, they predicted that about 37% of mutations in TFBSs were deleterious, about half the fraction estimated for 0-fold nonsynonymous sites in coding sequences.

The New Wave: Studies Based on Intraspecies Polymorphism

Divergence-based analyses, while informative, are fundamentally limited by the relatively long evolutionary time periods associated with the accumulation of fixed differences between species. Irregularities in the evolutionary process during these periods—for example, due to changes in the locations or boundaries of CREs, or changes in selective pressures—can weaken the signal of natural selection, causing its influence to be underestimated. This problem can be mitigated by working instead with data describing genetic variation within a single species [25]. Intraspecies polymorphism provides a window into much more recent evolutionary processes, on the time scale of genealogies of individuals rather than species phylogenies (for humans, roughly 1M years or less), during which the evolutionary process is likely to be more homogeneous. It has been demonstrated at numerous individual loci that patterns of human polymorphism can reveal the influence of natural selection on CREs [26–29].

Several groups have recently used this approach in genome-wide analyses of CREs, taking advantage of the abundant high-quality human polymorphism data now available. Because polymorphisms are sparse along the genome, these groups have generally pooled data across many similar loci. For example, Mu and colleagues examined human polymorphism data from the 1000 Genomes Project in various classes of coding and noncoding elements, including ChIP-seq-supported TFBSs [30]. The authors found that TFBSs were significantly constrained, but less so than coding sequences. Negative selection dominated in their tests, with no sign of pervasive positive selection. They observed stronger constraint in bound than in unbound TFBSs, in TFBSs proximal to transcription start sites (TSSs) than in ones distal to TSSs, and in TFBSs with strong rather than weak ChIP-seq signals. The related work of Khurana et al. further showed that mutations that decrease the matching score of a motif were enriched for rare alleles compared to ones that did not [31]. However, Khurana and colleagues found evidence of contributions from positive selection as well as negative selection in several types of regulatory elements, including DNase-I hypersensitive sites (DHSs) and sequence-specific TFBSs.

In another analysis of 1000 Genomes data, Ward and Kellis examined mean SNP density, heterozygosity, and derived allele frequency in various noncoding regions identified as having "biochemical activity" by the Encyclopedia of DNA Elements (ENCODE) project [32]. They observed significant constraint in putative regulatory regions identified by a wide variety of experimental assays. Interestingly, they found such evidence both for regions that were conserved across mammalian species and ones that were nonconserved, suggesting that a substantial fraction of functional noncoding elements reside outside of mammalian-conserved regions. In a similar study, Vernot et al. analyzed 53 high coverage individual genome sequences in more than 700 motifs within DHSs from 138 cell and tissue types, finding that many of these motifs were signficantly constrained [33].

A separate line of research has considered patterns of nucleotide diversity in flanking sequences of noncoding regions conserved across mammals, which are likely enriched for CREs [34–37]. These studies have come to conflicting conclusions, with some arguing for a prominent role for hitchhiking (HH) from positively selected sites in regulatory elements [34, 37], and others maintaining that the observed patterns are more consistent with background selection (BGS) from negative selection [35, 36]. More work is needed to resolve this controversy over the relative roles of positive and negative selection in shaping CREs.

A Fusion of the Old and the New: Joint Consideration of Divergence and Polymorphism

Population genomic data, too, has limitations when used as the sole source of information about natural selection. As noted above, it can be difficult to distinguish between positive and negative selection based on patterns of polymorphism alone (both forces reduce diversity; see Figure 1). Another major challenge is accounting for the effects of population bottlenecks, expansions, and other demographic processes, which can profoundly influence allele frequencies even in the absence of natural selection [38]. These problems can be alleviated by jointly considering intraspecies polymorphism and divergence from a neighboring species, an idea that has been used for decades in the analysis of protein-coding genes [39–41]. Classical approaches of this kind, such as the McDonald-Kreitman (MK) test [40], compare relative rates of polymorphism and divergence in putatively functional and nonfunctional (typically, nonsynonymous and synonymous) classes of sites. Under neutral drift, fixation should occur randomly for both classes of sites, causing the ratios of polymorphisms and fixed differences to be approximately equal. Departures from this neutral expectation provide information about natural selection (Figure 1).

An early attempt at a joint analysis of polymorphism and divergence of CREs, by Torgerson and colleagues, examined conserved noncoding regions flanking more than 15,000 protein-coding genes, using polymorphism data from 15 African Americans and 20 European Americans as well as the chimpanzee genome [42]. The authors made use of an extension of the MK test that permits estimation of selection coefficients [43], adapting it for use with noncoding sequences. Consistent with previous analyses, they found clear evidence of purifying selection in these regions. In addition, they found a significant excess of fixed differences relative to polymorphic sites, indicating positive selection on at least some CREs. In the study discussed above [24], Gaffney and colleagues also made limited use of polymorphism data, attempting to compute the fraction of fixed differences driven by positive selection (α) in CREs using a simple estimator based on the MK framework (see [41]). In contrast to Torgerson et al., they found no significant evidence of positive selection on CREs, but their power appeared to be quite weak.

Arbiza and colleagues attempted to address previous limitations in both models and data in a large-scale analysis of TFBSs based on ChIP-seq data from the ENCODE project [44]. Using a new probabilistic model and inference method called INSIGHT, the authors analyzed 1.4 million binding sites from 78 TFs, together with genetic variation data from the human, chimpanzee, orangutan, and rhesus macaque genome sequences, and 54 high-coverage human genome sequences. They found strong evidence of both positive and negative selection in TFBSs, with somewhat more positive selection, more weak negative selection, and less strong negative selection than in protein-coding genes. The authors estimated that, overall, there have been at least as many adaptive substitutions in CREs as in protein-coding genes since the human-chimpanzee divergence, consistent with King and Wilson's conjecture almost forty years earlier.

Another interesting observation from this study was that regulatory regions exhibited a large excess of weakly deleterious segregating mutations compared with protein-coding genes, suggesting considerable genetic load associated with gene regulation. This finding is concordant with a recent analysis of genetic association data, which found that regulation-associated DNase-I hypersensitivity sites accounted for almost 80% of the heritability for 11 common diseases [45]. Together, these findings suggest that a shift toward weaker negative selection in CREs may somewhat paradoxically result in an enrichment for heritable disease-causing segregating variants, because these variants are less efficiently eliminated by natural selection than those in protein-coding genes.

The Next Frontier

Most studies of *cis*-regulatory evolution in humans, including all of those discussed so far, have assumed that binding sites maintain stable positions at orthologous genomic locations over evolutionary time, and that fitness effects can be measured by patterns of variation at individual nucleotide positions. In reality, however, natural selection acts on nucleotides in TFBSs only indirectly, through the effects of those nucleotides on transcriptional output. These effects, in turn, occur through a complex and incompletely understood set of physical interactions involving multiple TFs and cofactors, the core transcriptional machinery, the DNA sequence, the local chromatin, and the surrounding aqueous environment [46, 47] (see Figure 2). Recognizing the full complexity of transcriptional regulation will be essential for a complete understanding of its evolution in humans and other species.

Biophysical Models of Binding-Site Evolution

A pioneering series of papers by Lässig and colleagues began to explore this complex intersection of biophysics and evolution using models that treated the free energy of TF binding to DNA as a quantitative phenotype, which served as the basis of an explicit fitness landscape. Evolutionary trajectories over this landscape were then considered [48–51] (see also [52]). Despite assuming an additive model for nucleotide-specific binding energies, the authors obtained highly nonlinear fitness landscapes, reflecting epistasis between regulatory nucleotides. In both prokaryotes and yeast, they found evidence for widespread compensatory mutations and relatively frequent gain and loss of binding sites.

Following these observations, Moses developed statistical tests for natural selection in terms of changes in predicted binding affinity resulting from single nucleotide changes under standard position-weight-matrix (PWM) models of binding [53]. Another study showed that evolutionary events tended to preserve binding affinity in Drosophila [54]. More recently, Bullaughey studied the evolution of enhancers by combining a thermodynamic sequence-to-expression model [55] with a Gaussian expression-to-fitness model [56]. His simulation study suggested strong interdependencies between nucleotides and an important role for neutral substitutions in changes to the functional organization of enhancers. Finally, in an analysis of well characterized *cis*-regulatory modules in Drosophila, He et al. found bulk evidence for positive selection contributing to both gain and loss of binding sites and for purifying selection maintaining existing TFBSs [57].

Another recent series of papers has focused on the development of improved biophysical models of TF binding to DNA, generally without consideration of evolution. A full review of this literature is outside the scope of the present article, but examples include models that consider combinatorial interactions among TFs [58–62], nucleosome positioning and/or chromatin accessibility [63–66], and the three-dimensional structure of DNA binding sites [67] (see [47] for a related review). More work is needed to consider the evolutionary implications of biophysical models of this type, but it seems likely that inferences of the distribution of fitness effects of regulatory mutations in humans will change significantly when richer, more realistic models of binding site structure and function are considered.

Improved Characterizations of Binding Affinity

Even the sophisticated biophysical models discussed in the previous section have tended to maintain the assumption of additive contributions of individual nucleotides to TF binding affinity, corresponding to an assumption of site independence in statistical motif models [68, 69]. This assumption appears to be adequate for most TFs, but numerous violations have been observed [70–72]. Nevertheless, statistical methods that attempt to recover the full correlation structure of TF binding preferences from sequences [71, 73, 74] have not been widely adopted.

These challenges have led to intense interest in harnessing high-throughput genomic technologies to produce direct measurements of binding affinity for all possible binding sites and large numbers of TFs. Widely variable strategies have been employed, including microwell-based assays [75, 76], protein-binding microarrays [67, 77–79], mechanically induced trapping of molecular interactions (MITOMI) [80], high-throughput systematic evolution of ligands by exponential enrichment (SELEX), [81–83], and, most recently, adaptation of the Illumina sequencing platform to directly measure binding affinities of proteins to DNA [84] (see [85] for a review as of 2010). In addition to finding further evidence of positional interdependence [79, 83, 84, 86, 87], studies based on these techniques have revealed, among other features, unexpected dimeric modes of binding [82], numerous TFs that recognize multiple sequence motifs [79], and important influences of sequences flanking core binding sites owing to their effects on DNA shape [67, 83]. However, the rich models of binding affinity enabled by these powerful technologies have yet to be integrated into evolutionary models.

Evolutionary Turnover of Cis-*Regulatory Elements*

As alluded to in the previous section, there is strong evidence that individual CREs in many species, including humans, are gained and lost over time, a phenomenon known generally as "turnover" [88–90]. Turnover of CREs has been extensively studied over the past decade [56, 57, 91–98] but, overall, it remains poorly understood. For example, it is still unclear how frequently turnover occurs overall, how much it varies across species, TFs, and genomic contexts, how commonly gains and losses are compensatory, and how all of these processes impact inferences of selection. Recent studies that make use of high-throughput functional genomic techniques applied uniformly across species [99, 100] have helped to shed additional light on turnover of CREs, but these studies also have limitations. For example, it is not clear how many of the assayed binding events directly influence gene expression, what role false negatives and false positives play in apparent differences, and in some cases sample sizes have been insufficient to distinguish within-species variation from between-species divergence. In our view, it will be essential to develop improved methods for integrating evolutionary and biophysical models with large-scale functional genomic data, to develop a more complete understanding of the complex processes by which CREs evolve.

Acknowledgments

A.S. is supported by NIH grants R01-GM102192 and R01-HG007070. L.A. is supported by NIH grant R01-HG006849 to Alon Keinan. The authors' thinking about transcriptional regulation has benefited from many valuable discussions with Andre Martins, Charles Danko, Leighton Core, and John Lis.

References and recommended reading

[1] Chimpanzee Sequencing and Analysis Consortium . Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005;437(7055):69–87.

[2] King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science 1975;188:107–16.

[3] Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science 1969;165(3891):349–57.

[4] Wilson AC, Maxson LR, Sarich VM. Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc Natl Acad Sci USA 1974;71(7):2843–7.

[5] Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 1961;3:318–56.

[6] Monod J, Jacob F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harb Symp Quant Biol 1961;26:389–401.

[7] Stern DL. Evolutionary developmental biology and the problem of variation. Evolution 2000;54(4):1079–91.

[8] Carroll SB. Evolution at two levels: on genes and form. PLoS Biol 2005;3(7):e245.

[9] Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 2007;8(3):206–16.

[10] Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution 2007;61(5):995–1016.

[11] Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 1977;267:275–6.

[12] Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 1985;2:150–74.

[13] Nei M, Gojobori T. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986;3:418–26.

[14] Keightley PD, Lercher MJ, Eyre-Walker A. Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol 2005;3(2):e42.

[15] Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res 2005;15(10):1373–8.

[16] Eory L, Halligan DL, Keightley PD. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol 2010;27(1):177–92.

[17] Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 1998;148:929–36.

[18] Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 2002;19(6):908–17.

[19] Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet 2007;39(9):1140–4.

[20] Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 2006;443(7108):167–72.

[21] Prabhakar S, Noonan JP, Paabo S, Rubin EM. Accelerated evolution of conserved noncoding sequences in humans. Science 2006;314(5800):786.

[22] Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, et al. Human-specific gain of function in a developmental enhancer. Science 2008;321(5894):1346–50.

[23] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110–21.

[24] Gaffney DJ, Blekhman R, Majewski J. Selective constraints in experimentally defined primate regulatory regions. PLoS Genet 2008;4(8):e1000157.

[25] Lawrie DS, Petrov DA. Comparative population genomics: power and principles for the inference of functionality. Trends Genet 2014;30(4):133–9.

[26] Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 2000;66(5):1669–79.

[27] Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 2002;70(2):369–83.

[28] Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol 2005;3:e387.

[29] Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 2007;39:31–40.

[30] Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res 2011;39(16):7058–76.

[31] Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 2013;342(6154):1235587.

[32] Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 2012;337(6102):1675–8.

[33] Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, et al. Personal and population genomics of human regulatory variation. Genome Res 2012;22(9):1689–97.

[34] Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet 2009;5(1):e1000336.

[35] McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 2009;5:e1000471.

[36] Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science 2011;331(6019):920–4.

[37] Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res 2014;.

[38] Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark A. Recent and ongoing selection in the human genome. Nat Rev Genet 2007;8:857–68.

[39] Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics 1987;116:153–9.

[40] McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991;351:652–4.

[41] Smith NG, Eyre-Walker A. Adaptive protein evolution in Drosophila. Nature 2002;415:1022–4.

[42] Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, et al. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 2009;5:e1000592.

[43] Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. The cost of inbreeding in Arabidopsis. Nature 2002;416:531–4.

[44] Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, et al. Genome-wide inference of natural selection on

human transcription factor binding sites. Nat Genet 2013;45(7):723–9.

[45] Gusev A, Lee SH, Neale BM, Trynka G, Vilhjalmsson BJ, Finucane H, et al. Regulatory variants explain much more heritability than coding variants across 11 common diseases. bioRxiv 2014;URL: http://dx.doi.org/10.1101/004309.

[46] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, et al. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 2003;20(9):1377–419.

[47] Siggers T, Gordan R. Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res 2014;42(4):2099–111.

[48] Berg J, Willmann S, Lässig M. Adaptive evolution of transcription factor binding sites. BMC Evol Biol 2004;4:42.

[49] Mustonen V, Lässig M. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc Natl Acad Sci USA 2005;102(44):15936–41.

[50] Lässig M. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics 2007;8 Suppl 6:S7.

[51] Mustonen V, Kinney J, Callan CG, Lässig M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proc Natl Acad Sci USA 2008;105(34):12376–81.

[52] Gerland U, Hwa T. On the selection and evolution of regulatory DNA motifs. J Mol Evol 2002;55(4):386–400.

[53] Moses AM. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol Biol 2009;9:286.

[54] Kim J, He X, Sinha S. Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet 2009;5(1):e1000330.

[55] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 2008;451(7178):535–40.

[56] Bullaughey K. Changes in selective effects over time facilitate turnover of enhancer sequences. Genetics 2011;187(2):567–82.

[57] He BZ, Holloway AK, Maerkl SJ, Kreitman M. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet 2011;7(4):e1002053.

[58] Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 2009;457(7226):215–8.

[59] Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci USA 2010;107(20):9158–63.

[60] Gertz J, Gerke JP, Cohen BA. Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions. Theor Popul Biol 2010;77(1):1–5.

[61] He X, Samee MA, Blatti C, Sinha S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol 2010;6(9).

[62] Cheng Q, Kazemian M, Pham H, Blatti C, Celniker SE, Wolfe SA, et al. Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. PLoS Genet 2013;9(8):e1003571.

[63] Raveh-Sadka T, Levo M, Segal E. Incorporating nucleosomes into thermodynamic models of transcription regulation. Genome Res 2009;19(8):1480–96.

[64] Kaplan T, Li XY, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, et al. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet 2011;7(2):e1001290.

[65] Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol 2011;12(4):R34.

[66] Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet 2012;44(7):743–50.

[67] Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep 2013;3(4):1093–104.

[68] Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol 1987;193(4):723–50.

[69] Stormo GD. DNA binding sites: representation and discovery. Bioinformatics 2000;16(1):16–23.

[70] Bulyk ML, Johnson PL, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 2002;30(5):1255–61.

[71] Zhou Q, Liu JS. Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 2004;20(6):909–16.

[72] Guertin MJ, Martins AL, Siepel A, Lis JT. Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet 2012;8(3):e1002610.

[73] Barash Y, Elidan G, Kaplan T, Friedman N. Modeling dependencies in protein-dna binding sites. In: Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB). 2003, p. 28–37.

[74] Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, et al. Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics 2005;21(11):2657–66.

[75] Hallikas O, Taipale J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. Nat Protoc 2006;1(1):215–22.

[76] Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 2006;124(1):47–59.

[77] Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 2006;24(11):1429–35.

[78] Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 2008;133(7):1266–76.

[79] Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science 2009;324(5935):1720–3.

[80] Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. Science 2007;315(5809):233–7.

[81] Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. PLoS Comput Biol 2009;5(12):e1000590.

[82] Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res 2010;20(6):861–73.

[83] Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell 2013;152(1-2):327–39.

[84] Nutiu R, Friedman RC, Luo S, Khrebtukova I, Silva D, Li R, et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat Biotechnol 2011;29(7):659–64.

[85] Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev Genet 2010;11(11):751–60.

[86] Zhao Y, Ruan S, Pandey M, Stormo GD. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics 2012;191(3):781–90.

[87] Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol 2013;31(2):126–34.

[88] Ludwig MZ, Patel NH, Kreitman M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 1998;125(5):949–58.

[89] Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 2000;403(6769):564–7.

[90] Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 2002;19(7):1114–21.

[91] Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2006;2(10):e130.

[92] Doniger SW, Fay JC. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 2007;3(5):e99.

[93] Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 2007;39:730–2.

[94] Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol 2010;8(3):e1000343.

[95] Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 2010;328(5981):1036–40.

[96] Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in transcription factor binding among humans. Science 2010;328:232–5.

[97] Weirauch MT, Hughes TR. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet 2010;26(2):66–74.

[98] Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet 2014;15(4):221–33.

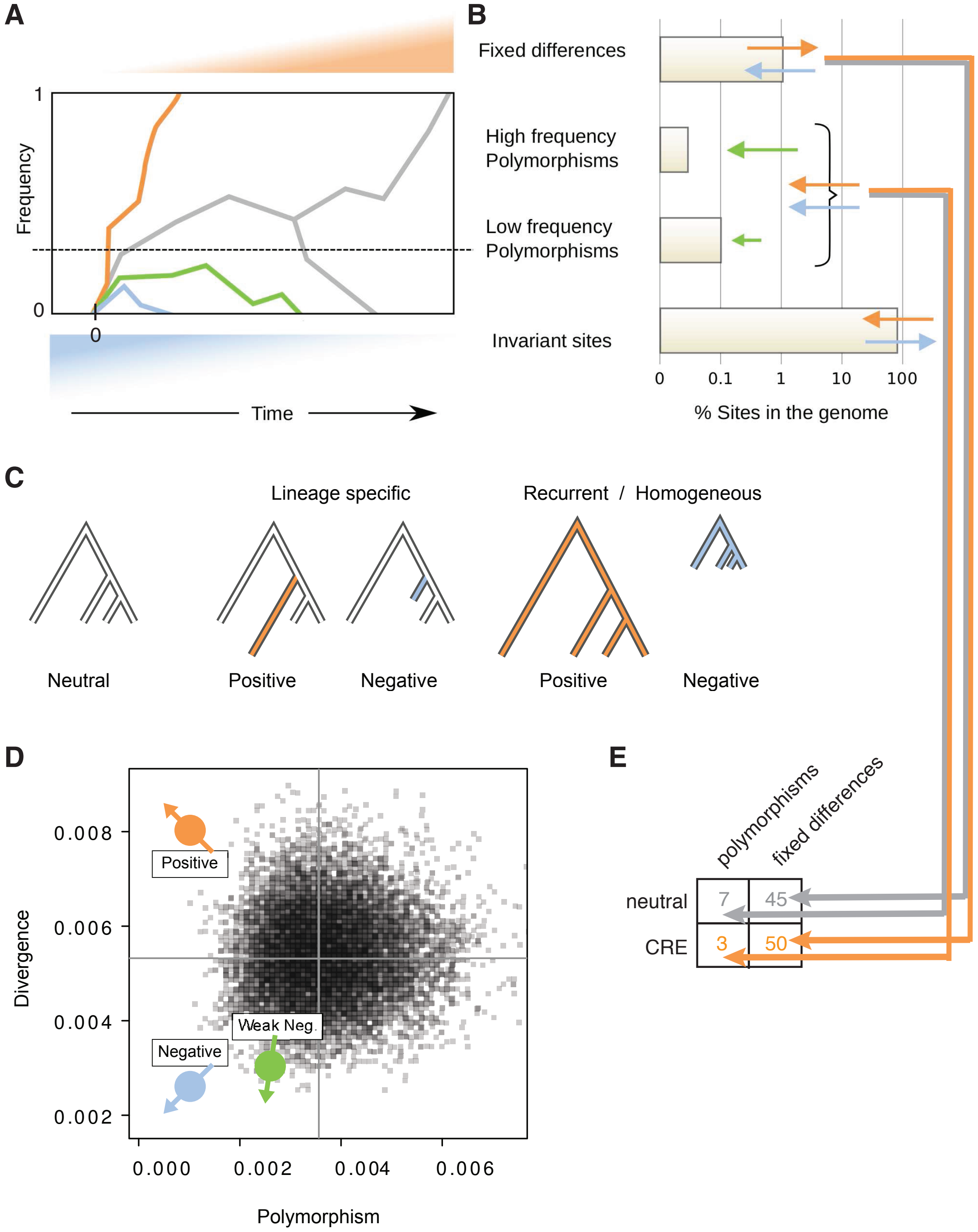
[99] Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet 2012;8(6):e1002789.

[100] Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 2013;154(3):530–40.

Figure Legends

Figure 1: (A) Frequency as a function of time for hypothetical mutations experiencing neutral drift (gray), weak negative (green), strong negative (blue), or positive (orange) selection. The plot assumes a new mutation occurs in a single individual in the population at time 0. Neutral drift typically causes mutations to be lost (lower gray fork) but occasionally drives them to fixation (upper gray fork). Negative selection essentially guarantees eventual loss, but if it is sufficiently weak (green plot), mutations may segregate at low frequencies for some time. Positive selection (orange plot) causes mutations to reach fixation at higher rates than neutral drift. Notice that the time until fixation or loss is substantially reduced for mutations under strong selection (positive or negative), implying that they are unlikely to be observed in a polymorphic state. (B) Steady-state numbers of invariant sites, low frequency (derived allele) polymorphisms, high frequency polymorphisms, and fixed differences under neutral drift, expressed as hypothetical percentages of nucleotide sites. These represent equilibrium frequencies for the process depicted in panel (A) for a given divergence time, assuming a steady flow of new mutations. Positive selection (orange arrows) increases fixed differences, reduces invariant sites, and reduces polymorphisms. Strong negative selection (blue arrows) reduces fixed differences and polymorphisms and increases invariant sites. Weak negative selection (green arrows) is similar but allows some low frequency polymorphisms to remain. (C) Phylogenies with branch lengths proportional to rates at which fixed differences occur along lineages. Positive or negative selection can be identified by significant increases or decreases, respectively, in the fixation rates relative to the neutral expectation. Different likelihood ratio tests can identify lineage-specific or recurrent/homogeneous selective pressures. (D) Scatter plot of polymorphism vs. divergence rates under neutral drift, generated by simulations based on parameters reflecting real human populations [44] (black points). Colored points show hypothetical positions of sequences under positive (orange), strong negative (blue), and weak negative (green) selection. Notice that positive and negative selection are distinguishable by their joint effects on polymorphism and divergence rates, but not by polymorphism rates alone. (E) 2×2 contingency table used for McDonald-Kreitman (MK) test for selection on a *cis*-regulatory element (CRE). The test evaluates the probability of the observed data under the null hypothesis that the relative polymorphism and divergence counts are independent of the labels "neutral" and "CRE". The classes of sites are chosen to be similar to one another to avoid potential biases from mutation rate variation and demography. Rejection of the null hypothesis therefore implies a departure from the neutral expectation of equal fixation rates. Note the connections with the visual representations used in panels (B) and (D). The MK test can be thought of as comparing the relative heights of the first bar and the next two bars combined in panel (B), for neutral vs. CRE sites (see arrows). It can also be thought of as testing for extreme departures from a diagonal line in panel (D) running through the neutral points from bottom left to top right. In this case, the counts reflect an excess of fixed differences in the CRE, suggesting positive selection. Notice that strong negative selection is not a problem for the MK test, because it reduces the effective mutation rate, but weak negative selection can bias the test by partially canceling the effects of positive selection.

Figure 2: Some of the many factors that may influence the evolution of *cis*-regulatory elements.



	polymorphisms	fixed differences
neutral	7	45
CRE	3	50

Figure 1:

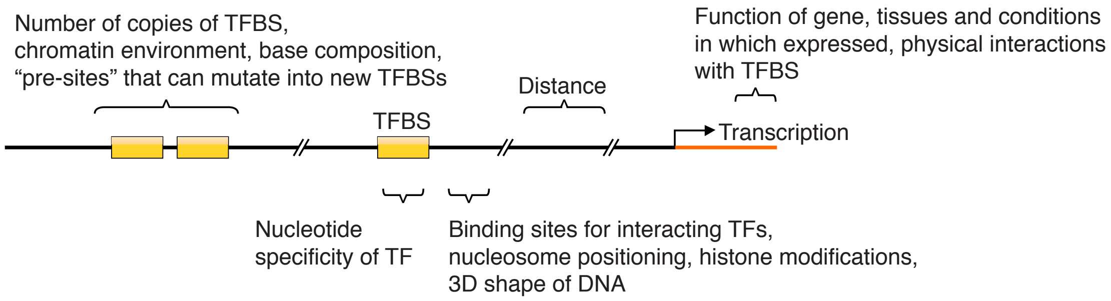

Number of copies of TFBS,
chromatin environment, base composition,
“pre-sites” that can mutate into new TFBSs
Function of gene, tissues and conditions
in which expressed, physical interactions
with TFBS
Distance
TFBS
Transcription
Nucleotide
specificity of TF
Binding sites for interacting TFs,
nucleosome positioning, histone modifications,
3D shape of DNA

Figure 2: