Review Article

The Evolution of Transcriptional Regulation in Eukaryotes

Gregory A. Wray, Matthew W. Hahn, Ehab Abouheif, James P. Balhoff, Margaret Pizer, Matthew V. Rockman, and Laura A. Romano

Department of Biology, Duke University

Gene expression is central to the genotype-phenotype relationship in all organisms, and it is an important component of the genetic basis for evolutionary change in diverse aspects of phenotype. However, the evolution of transcriptional regulation remains understudied and poorly understood. Here we review the evolutionary dynamics of promoter, or *cis*-regulatory, sequences and the evolutionary mechanisms that shape them. Existing evidence indicates that populations harbor extensive genetic variation in promoter sequences, that a substantial fraction of this variation has consequences for both biochemical and organismal phenotype, and that some of this functional variation is sorted by selection. As with protein-coding sequences, rates and patterns of promoter sequence evolution differ considerably among loci and among clades for reasons that are not well understood. Studying the evolution of transcriptional regulation poses empirical and conceptual challenges beyond those typically encountered in analyses of coding sequence evolution: promoter organization is much less regular than that of coding sequences, and sequences required for the transcription of each locus reside at multiple other loci in the genome. Because of the strong context-dependence of transcriptional regulation, sequence inspection alone provides limited information about promoter function. Understanding the functional consequences of sequence differences among promoters generally requires biochemical and in vivo functional assays. Despite these challenges, important insights have already been gained into the evolution of transcriptional regulation, and the pace of discovery is accelerating.

1 Introduction

A gene embedded in random DNA is inert. In the absence of sequence motifs and proteins capable of directing transcription, the protein it encodes will remain invisible to selection. Every gene with a phenotypic impact is flanked by regulatory sequences that, in conjunction with the expression and activity of proteins encoded elsewhere, regulate when expression occurs, at what level, under what environmental conditions, and in which cells or tissues. Transcriptional regulatory sequences are as important for gene function as the coding sequences that determine the linear array of amino acids in a protein.

Transcriptional regulation is also a crucial contributor to evolutionary change in the genotype-phenotype relationship. Understanding the dynamic link between genotype and phenotype remains a central challenge in evolutionary biology (Wright 1982; Raff 1996; Wilkins 2002). Enormous advances have been made during the past few decades in understanding the dynamics of alleles within populations, the role of genes during development, and the evolution of phenotype. Although these studies have progressed along nearly independent paths (for historical perspectives, see Raff [1996] and Wilkins [2002]), they have recently begun to intersect in fruitful and exciting ways in studies of gene expression. This work is making substantial contributions to the understanding of how the genotype-phenotype relationship evolves.

The goal of this review is to bring transcriptional regulation into the mainstream of molecular evolution. We are concerned here with promoters (*cis*-regulatory sequences that influence transcription) and transcription factors (proteins that interact with these sequences). Throughout, we emphasize three general points. First, changes in transcriptional regulation comprise a quantitatively and qualitatively significant component of the genetic basis for evolutionary change. Second, understanding how transcriptional regulation evolves requires a clear grasp of how the relevant macromolecules interact and function in living cells. And third, studying the evolution of transcriptional regulation poses unique and significant challenges to both empirical and analytical approaches. These challenges are balanced, however, by extraordinary opportunities to extend and deepen our understanding of the genetic basis for phenotypic evolution.

Key words: binding site, enhancer, evolution of development, genotype-phenotype relationship, promoter, transcription factor.

E-mail: gwray@duke.edu.

Mol. Biol. Evol. 20(9):1377–1419. 2003
DOI: 10.1093/molbev/msg140
Molecular Biology and Evolution, Vol. 20, No. 9,

2 Why Promoter Evolution Matters

Several recent reviews have argued that changes in transcriptional regulation constitute a major component of the genetic basis for phenotypic evolution (Doebley and Lukens 1998; Carroll 2000; Stern 2000; Tautz 2000; Theissen et al. 2000; Purugganan 2000; Wray and Lowe 2000; Carroll, Grenier, and Weatherbee 2001; Davidson 2001; Wilkins 2002). Although the authors reached similar conclusions, they provided limited evidence to support the claim that mutations affecting transcriptional regulation have important evolutionary consequences. In this section we therefore review the theoretical arguments and empirical evidence that transcriptional regulation plays a pervasive and important role in evolution.

2.1 Theoretical Arguments: Why Promoters Ought to Contribute to Phenotypic Evolution

Before direct evidence was available, a few far-sighted biologists argued on the basis of first principles that changes in gene expression should constitute an important part of the genetic basis for phenotypic change (Jacob and Monod

1961; Wallace 1963; Zuckerkandl 1963; Britten and Davidson 1969, 1971; King and Wilson 1975; Wilson 1975; Jacob 1977; Raff and Kaufman 1983). Their arguments were based in part on the realization that the phenotypic impact of a gene is a function of two distinct components: the biochemical activity of the protein it encodes and the specific conditions under which that protein is expressed and is therefore able to exert its activity. During subsequent decades, the field of molecular evolution focused on the evolutionary implications of the first component of function, while developmental biologists were more concerned with the functional implications of the second. The revival of ''evo-devo'' has focused attention on a more integrative view that encompasses both protein function and gene expression (Raff 1996; Wilkins 2002).

Four additional considerations suggest that transcriptional regulation ought to be evolutionarily important. (1) *Significant phenotypes*. Many authors have commented on the direct relationship between when or where a gene is expressed and the functionally significant phenotypes that might result from changing these parameters (Raff and Kaufman 1983; Gerhart and Kirschner 1997; Carroll 2000; Davidson 2001; Wilkins 2002). For instance, earlier expression of a hormone might result in accelerated growth, whereas ectopic expression of a transcription factor might result in a duplicated structure. Importantly, these phenotypic transformations can be independent of changes in protein sequences. Changes in precisely how transcription is regulated can also have significant phenotypic consequences (Paigen 1989). For instance, synthesizing a digestive enzyme in response to feeding or resource availability might prove advantageous compared with continuous production (Jacob and Monod 1961). Such changes may form the basis of polyphenism and phenotypic plasticity (Schlichting and Pigliucci 1998; Gilbert 2001). (2) *Coordinated pleiotropy*. Because the proteins that regulate transcription interact with batteries of functionally related genes, a mutation affecting the function or expression of a transcription factor can potentially produce a coordinated phenotypic response (Raff and Kaufman 1983; Gerhart and Kirschner 1997; Carroll, Grenier, and Weatherbee, 2001; Wilkins 2002). Mutations in the expression of transcriptional regulators are therefore not simply more pleiotropic, they are more likely to produce functionally integrated phenotypic consequences. (3) *The ''*Hox *paradox.''* The discovery that many developmental regulatory genes and their expression profiles are phylogenetically widespread within the plant and animal kingdoms (Gerhart and Kirschner 1997; Carroll et al. 2001) raises an obvious problem: How do orthologous regulatory proteins pattern anatomically disparate organisms? At least part of the answer seems to lie in evolutionary reorganization of gene networks, such that many interactions between these proteins and the collection of genes that they regulate has changed since flies and mice last shared a common ancestor (Wray and Lowe 2000; Davidson 2001; Wilkins 2002). (4) *Evolvability*. Promoters may be more ''evolvable'' than coding regions (Gerhart and Kirschner 1997; Stern 2000; Carroll, Grenier, and Weatherbee 2001; Wilkins 2002). Many promoters are organized into functional modules, each of which produces a discrete aspect of the overall expression profile (Arnone and Davidson 1997), confining pleiotropy and allowing selection to modify discrete aspects of the overall expression profile independently. In addition, many promoter alleles are likely to be codominant and thus immediately visible to selection, increasing the efficiency with which beneficial alleles are fixed and deleterious ones are eliminated.

2.2 Mutations in Transcriptional Regulation Influence Phenotype

Transcriptional regulation is an integral component of the way genotype is converted into phenotype. Many mutants that have emerged from genetic screens for developmentally important genes involve defects in transcriptional regulation (Wilkins 1993, 2002; Gilbert 2000). The four-winged fly that results from certain mutations in *Ubx* in Drosophila is perhaps the most famous: some mutations located in regulatory sequences affect the transcription profile, and others locating in exons alter the function of the protein in regulating the transcription of other genes (Bender et al. 1983; Simon et al. 1990). The phenotypic consequences of some *Ubx* promoter mutations are so distinct that they were originally thought to represent separate genes (Lewis 1978).

Numerous studies have documented correlations between gene expression and anatomy. (1) *Induced mutations*. The phenotypes of some induced mutations mimic natural differences between species. Examples include homeotic mutations in *Drosophila melanogaster* that mimic segment and appendage number and identity characteristic of other insects (Raff and Kaufman 1983; Carroll 1995), mutations in *Arabidopsis thaliana* and *Antirrhinum majus* that mimic the floral anatomy of other angiosperms (Lawton-Rauh et al. 2000), and mutations in *Caenorhabditis elegans* that mimic the tail anatomy of other nematodes (Fitch 1997). Because most of these induced mutations generally do not replicate the genetic basis for natural phenotypic differences (Carroll 1995; Budd 1999), however, convincing evidence of the evolutionary significance of changes in transcriptional regulation must come from natural cases. (2) *Comparisons of expression*. In many cases, a gene required for the development of a trait in one species shows a difference in expression in other species that correlates with a difference in that trait (e.g., Burke et al. 1995; Brakefield et al. 1996; Dudareva et al. 1996; Sinha and Kellogg 1996; Averof and Patel 1997; Stockhaus et al. 1997; Abzhanov and Kaufman 2000; Kopp et al. 2000; Yamamoto and Jeffery 2000; Beldade, Brakefield, and Long 2002; Bharathan et al. 2002; Hariri et al. 2002). A causal relationship is plausible but not proven in these cases, because comparisons of gene expression cannot by themselves demonstrate that a change in transcriptional regulation is the genetic basis for a phenotypic difference. (3) *Quantitative genetics*. Anatomical changes that accompanied the domestication of maize from teosinte are due in part to changes within the inferred promoter region of a single gene encoding the transcription factor teosinte-branched (Wang et al. 1999). Although this is a case of artificial selection, it involved natural (rather than induced)

genetic variation. Some differences in bristle patterns among Drosophila species are attributable to changes in promoter sequences (Stern 1998; Skaer and Simpson 2000; Sucena and Stern 2000). In other cases, genetic variation in gene expression levels shows strong associations with specific organismal phenotypes (Gerber, Fabre, and Planchon, 2000; Karp et al. 2000; Beldade, Brakefield, and Long 2002). Unfortunately, because of the confounding effects of linkage disequilibrium, quantitative genetics generally lacks the resolution to identify precise sequence differences that are responsible for particular phenotypes. When combined with experimental tests or case associations, however, specific sequence variants can be identified (Cooper 1999). Using this approach, more than 160 segregating promoter variants that influence transcription have been identified in humans (Cooper 1999; Rockman and Wray 2002), and several have been identified in *Drosophila melanogaster* (e.g., Robin et al. 2002).

2.3 Natural Populations Harbor Considerable Functional Variation in Gene Expression

Many examples of variation in gene expression are known from natural populations. (1) *Spatial extent of expression*. In rainbow trout, an allele of PGM1 conferring expression in the liver is associated with faster prehatching growth (Allendorf, Knudsen, and Phelps 1982; Allendorf, Knudsen, and Leary 1983). The spatial expression of amylase in the midgut varies within both *Drosophila melanogaster* and *D. pseudoobscura*; the genetic basis in both cases is *trans* and responds to artificial selection in *D. pseudoobscura* (Abraham and Doane 1978; Powell 1979; Powell and Licthenfels 1979). The spatial extent of expression of the transcription factor Distal-less within the wing of the butterfly *Bicyclus anynana* varies in correlation with wing color pattern, and it also responds to artificial selection (Beldade, Brakefield, and Long 2002). (2) *Level of expression*. Intraspecific differences in expression have been noted for GPDH in both larvae and adults of *D. melanogaster* (Laurie-Ahlberg and Bewley 1983); β-*glucuronidase* in *Mus domesticus* (Pfister et al. 1982; Bush and Paigen 1992); *Cyp6g1*, a *cytochrome P450* family gene, in *D. melanogaster* (Daborn et al. 2002); and *prolactin* in the teleost *Oreochromis niloticus* (Streelman and Kocher 2002). In all four cases, most or all of the polymorphisms described are in *cis*. Many additional examples are known from humans, where nearly two-thirds of the known functional polymorphisms in *cis*-regulatory sequences have a greater than twofold impact on transcription rates (Rockman and Wray 2002). (3) *Inducibility of expression*. Inducibility of amylase expression in response to a starch diet varies within *D. melanogaster* and responds to artificial selection (Matsuo and Yamazaki 1984; Klarenberg, Sikkema, and Scharloo 1987); expression of β-*glucuronidase* in response to androgen varies within *Mus domesticus* (Bush and Paigen 1992); and three different mobile element insertions into the promoter of *hsp70* reduce transcription in response to thermal stress in *D. melanogaster* populations (Lerman et al. 2003). Several other examples of variation in inducibility are known from humans (Rockman and Wray 2002). In the human and *hsp70* cases, the genetic basis is known to reside in *cis*.

Additional studies have estimated the extent of heritable genetic variation in gene expression within populations. (1) *Protein-based surveys*. Several studies have measured levels of variation in gene expression from 1- or 2-dimensional protein gels in a variety of organisms: *Zea mays* (Burstin et al. 1994; Damerval et al. 1994; de Vienne et al. 2001), *Pinus pinaster* (Costa and Plomion 1999), *Glycine max* (Gerber, Fabre, and Planchon 2000), *Mus musculus* (Klose et al. 2002), and *Homo sapiens* (Enard et al. 2002*a*). Studies with the first three organisms documented that protein abundance has a strong genetic component, and all of these studies found that populations contain considerable variation in expression level for most of the proteins surveyed. In *D. melanogaster*, chromosome substitution lines show substantial levels of variation in gene expression as measured by enzyme activities (Laurie-Ahlberg et al. 1980; Wilton et al. 1982; Clark 1990). Although protein abundance and enzyme activity are indirect indices of transcription, these results suggest considerable genetic variation for gene expression in general. (2) *mRNA-based surveys*. More direct estimates of variation in transcription come from microarray analyses that survey thousands of loci. Studies in mice (Karp et al. 2000; Schadt et al. 2003), humans (Schadt et al. 2003), the teleost *Fundulus heteroclitus* (Oleksiak, Churchill, and Crawford 2002), *D. melanogaster* (Jin et al. 2001; Rifkin, Kim, and White 2003), *Zea mays* (Schadt et al. 2003), and *Saccharomyces cerevisiae* (Cavalieri, Townsend, and Hartl 2000; Brem et al. 2002), all indicate that genetic variation in transcript abundance is pervasive within populations. Much of this variation may be heritable. Schadt et al. (2003) found that 33% of the 23,574 loci surveyed from a cross of two inbred strains of mice showed a genetic component for expression differences within the liver, 29% of the 2,726 loci surveyed from 56 humans belonging to four families showed a heritable difference in expression within lymphoblasts, and 18,805 genes consistently differed in transcription within ear leaf tissue among progeny from a cross of two maize strains. What proportion of the genetic basis for this variation resides in the promoters of the genes showing transcriptional variation (*cis*) or in the sequences or expression profiles of their upstream regulators (*trans*) has been examined in a few cases. Quantitative trait loci (QTL) underlying variation in expression of at least 32% of 570 variably expressed transcripts in yeast mapped in *cis* (Brem et al. 2002), whereas the comparable fraction of genes with *cis*-acting QTL in mouse liver is even higher (Schadt et al. 2003). Reverse transcriptase polymerase chain reaction (RT-PCR) offers more reliable quantitation than microarrays, and it also provides a means of directly comparing transcription rates among alleles. In a preliminary survey of 69 loci in four inbred lines of *Mus musculus*, Cowles et al. (2002) found quantitative and tissue-specific variation among alleles at 4 loci. Using a similar approach, Yan et al. (2002) found evidence of variation in gene expression at 6 of 13 loci examined in humans. Taken together, microarray and RT-PCR surveys

of mRNA levels provide solid evidence of abundant genetic variation in transcriptional regulation in diverse species, and they suggest that much of this variation resides in *cis* regulatory sequences. (3) *Detailed analyses of promoter function*. The most extensive direct evidence of functional variation in promoter sequences now available comes from humans, where many specific polymorphisms have been identified through direct functional studies (Cooper 1999). Although the human genome is not particularly polymorphic, a typical individual is estimated to be heterozygous for a functional promoter polymorphism at ~40% of all loci (Rockman and Wray 2002). Comparable data do not yet exist for other species, but RT-PCR surveys (Cowles et al. 2002; Yan et al. 2002) provide a rapid means of estimating heterozygosity that affects transcription at many loci.

2.4 Natural Selection Operates on Allelic Variation in Promoters

Evidence for natural selection on eukaryotic promoter alleles comes from a variety of sources (also see section 4.7). (1) *Human populations*. Promoter polymorphisms at numerous loci in humans have functional consequences that influence diverse aspects of physiology, behavior, anatomy, and life history (Cooper 1999; Rockman and Wray 2002). Some of these promoter alleles have likely fitness consequences (for examples, see next paragraph and section 4.7). (2) *Wild populations*. A latitudinal cline of LDH promoter allele frequencies in the teleost *Fundulus heteroclitus* is probably maintained by temperature differences (Crawford, Segal, and Barnett 1999; Segal, Barnett, and Crawford 1999). Two other cases, mentioned earlier, are known from *D. melanogaster*: promoter alleles segregating at both *Cyp6G1* and *hsp70* appear to be under selection in wild populations (Daborn et al. 2002; Lerman et al. 2003). (3) *Artificial selection and experimental evolution*. Domestication of maize involved selection on the inferred regulatory region of the *tb* locus (Wang et al. 1999). Studies with yeast point to regulation of transcription as a critical component of adaptive change. Adaptation of *Saccharomyces cerevisiae* to glucose limitation was accompanied by twofold or greater changes in the abundance of transcripts from nearly 10% of all genes, consistently across replicates (Ferea et al. 1999). The evolution of drug resistance in experimental populations of *Candida albicans* correlated with overexpression of the four known resistance genes (Cowen et al. 2000). (4) *Sequence comparisons*. More extensive, but less direct, evidence that natural selection acts on promoters comes from cases of apparent evolutionary conservation of *cis*-regulatory sequences among distantly related species (for examples, see section 4.1). Consistent underrepresentation of specific sequence motifs provides evidence for genome-wide selection to remove spurious transcription initiation sequences in a broad diversity of prokaryotes (Hahn, Stajich, and Wray 2003).

Several examples of natural selection operating on transcriptional regulation involve pathogen-host interactions. For instance, some promoter alleles in *Mycobacterium tuberculanum* and hepatitis B alter transcription to the pathogen's benefit and may be under positive selection (Buckwold et al. 1997; Rinder et al. 1998; Lee et al. 2000; Kajiya et al. 2001). The origin and subsequent fixation of these mutations in separate host individuals demonstrates the ability of positive selection to operate in a predictable way on genetic variation within a promoter. Specific variants within the human immunodeficiency virus (HIV) promoter, including gains of binding sites for host nuclear factor kappa-B (NF-kB) and upstream stimulatory factor (USF), as well as functional modifications in the basal promoter, cause differences in the level of viral transcription (Montano et al. 1997; Jeeninga et al. 2000). The E subtype of HIV has significantly increased transcription rates and has gone to near fixation locally in southern Africa; it is associated with increased levels of secondary infections and may be under positive selection to the pathogens' advantage (Montano et al. 2000; Hunt, Johnson, and Tiemesse 2001). Conversely, human populations harbor promoter variants that influence susceptibility to pathogens or disease progression after infection. Because human generation times are much longer than those of pathogens, signatures of selection are more difficult to detect. Nonetheless, promoter allelles at *TNF*α, *IL-4*, *IL-10*, *FY*, *CCR5*, and *TGF*β influence mortality from a variety of viral, bacterial, and protoctistan pathogens and are likely to be under selection (Tournamille et al. 1995; Hamblin and Di Rienzo 2000; Shin et al. 2000; Thurz 2001; Bamshad et al. 2002; Meyer et al. 2002; Nakayama et al. 2002; Vidigal, Gemner, and Zein 2002). Some promoter alleles confer protection from one pathogen while increasing susceptibility to another (e.g., *TNF*α–380*A*: Meyer et al. 2002), raising the possibility of balanced polymorphisms.

2.5 Divergence in Promoter Function May Contribute to Reproductive Isolation

Changes in transcriptional regulation may also be important in speciation. The Dobzhansky-Muller model of speciation requires interspecific differences at pairs of interacting loci (Dobzhansky 1936; Muller 1942). Because of the large number of highly specific interactions that occur between proteins and DNA within promoters, these regions represent likely sites for postzygotic isolation resulting from multilocus epistasis (Johnson and Porter 2000). Empirical support comes from genetic loci that are involved in reproductive isolation. Only four such loci have been identified definitively, and all have turned out to involve changes in transcriptional regulation: the coding sequence of the transcription factor Odysseus within the genus Drosophila (Ting et al. 1998); promoter sequences of *Xmk2* and *CKDN2X* within the teleost genus *Xiphophorus* (reviewed in Orr and Presgraves 2000); and a promoter polymorphism in *desaturase 2* of *D. melanogaster* that is correlated with intraspecific differences in mating behavior and may be involved in premating isolation (Fang, Takahashi, and Wu 2002).

3 Transcriptional Regulation in Eukaryotes

The familiar regularities that characterize coding sequences, in particular the genetic code, are absent from

promoters. Understanding the functional consequences of evolutionary differences in promoter sequences therefore requires a clear knowledge of the mechanisms of transcriptional regulation. In this section, we review the structure and function of eukaryotic promoters. The literature on this topic is vast, and the emphasis here is on features directly pertinent to promoter evolution. Our focus is on the transcription of protein-coding loci, which comprise the majority of genes in eukaryotic genomes and about which the most information is available. Transcriptional regulation in Eubacteria is distinct in many ways (Struhl 1999; Lewin 2000), whereas in Archaea it is not particularly well understood (although the latter shares many features with eukaryotic regulation: Bell and Jackson 1998; Weinzierl 1999). Neither prokaryotic group is covered in this review. For more detailed reviews of mechanisms of eukaryotic transcriptional regulation see Latchman (1998), Weinzierl (1999); Carey and Smale (2000), Lee and Young (2000), Lewin (2000), Davidson (2001), Locker (2001), and White (2001).

3.1 Promoters and Gene Expression

Only some of the genes in a eukaryotic cell are expressed at any given moment. The proportion and composition of transcribed genes changes considerably during the life cycle, among cell types, and in response to fluctuating physiological and environmental conditions (e.g., White et al. 1999; Iyer et al. 2001; Kayo et al. 2001; Mody et al. 2001; Arbeitman et al. 2002). Given that eukaryotic genomes contain on the order of 0.5 to 5×10^4 genes, regulating this differential gene expression requires an exceptionally complex array of specific physical interactions among macromolecules.

3.1.1 Most Regulation of Gene Expression Occurs at the Level of Transcription

Eukaryotes employ diverse mechanisms to regulate gene expression, including chromatin condensation, DNA methylation, transcriptional initiation, alternative splicing of RNA, mRNA stability, translational controls, several forms of post-translational modification, intracellular trafficking, and protein degradation (Lewin 2000; Alberts et al. 2002). Of these broad categories, the most common point of control is the rate of transcriptional initiation (Latchman 1998; Carey and Smale 2000; Lemon and Tjian 2000; White 2001). For virtually every eukaryotic gene where relevant information exists, transcriptional initiation appears to be the primary determinant, or one of the most important determinants, of the overall gene expression profile.

3.1.2 Transcriptional Regulation Is Primarily Gene-Specific

To a first approximation, the transcription of each gene in a eukaryotic genome is controlled independently. Operons (multi-locus transcripts regulated by a single promoter) are unusual in eukaryotes, a contrast with most prokaryotes. (Eukaryotic exceptions include the protozoan *Trypanosoma brucei* and the nematode *Caenorhabditis elegans*, where a substantial fraction of genes are transcribed as polycistronic mRNAs: Blumenthal 1998). Even paralogs within gene families are typically regulated independently and often have quite different expression profiles (e.g., Ferris and Whitt 1979; Fang and Brandhorst 1996; Christophides et al. 2000; Gu et al. 2002). Although a regulatory region sometimes directly influences the transcription of two loci (for examples, see section 3.3.7 and fig. 2), such cases apparently are uncommon. Distributed transcriptional regulation allows selection to fine-tune the expression profile of each gene independently.

3.1.3 Gene Expression Profiles Are Complex

Most genes are differentially transcribed across the life cycle, according to environmental conditions, in different cell types and regions, and among sexes. Transcriptional regulation is a highly dynamic process: rates of RNA synthesis can fluctuate by orders of magnitude, change over time scales of minutes, and differ among adjacent cells. Most genes have spatially and temporally heterogeneous expression profiles. Genes encoding regulatory proteins possess some of the most complex expression profiles. In metazoans and metaphytes, most such genes are expressed in several distinct domains (Gerhart and Kirschner 1997; Davidson 2001). For instance, the transcription factor Pax-6 is expressed at different times and at different levels in the telencephalon, hindbrain, and spinal cord of the central nervous system; in the lens, cornea, neural and pigmented retina, lacrimal gland, and conjunctiva of the eye; and in the pancreas (Kammandel et al. 1999). Where data are available, they link distinct phases of these complex expression profiles to distinct regulatory functions (Wray and Lowe 2000; Davidson 2001; Wilkins 2002). Although the transcription profiles of ''housekeeping'' genes are generally much simpler, most are transcribed at different levels among cell types and are shut down in response to extreme environmental conditions such as heat shock.

3.1.4 Promoters Integrate Information and Alter Transcription Accordingly

At its most fundamental level, the function of a promoter is to integrate information about the status of the cell in which it resides, and to alter the rate of transcriptional initiation of a single gene accordingly. The inputs that a promoter integrates can take many forms. The promoters of genes expressed during early development integrate spatial and temporal inputs to produce highly dynamic patterns of transcription in specific regions of the embryo (Davidson 2001; Wilkins 2002). The promoters of genes encoding housekeeping proteins are constitutively active, but they can shut down in response to specific conditions, such as heat shock or starvation (Pirkkala, Nykanen, and Sistonen 2001). Other promoters are off by default, but they can be activated in response to specific hormonal, physiological, or environmental cues (Benecke, Gaudon, and Gronemeyer 2001; Shore and Sharrocks 2001). These diverse inputs eventually reach promoters in the form of transcription factors, proteins that bind in

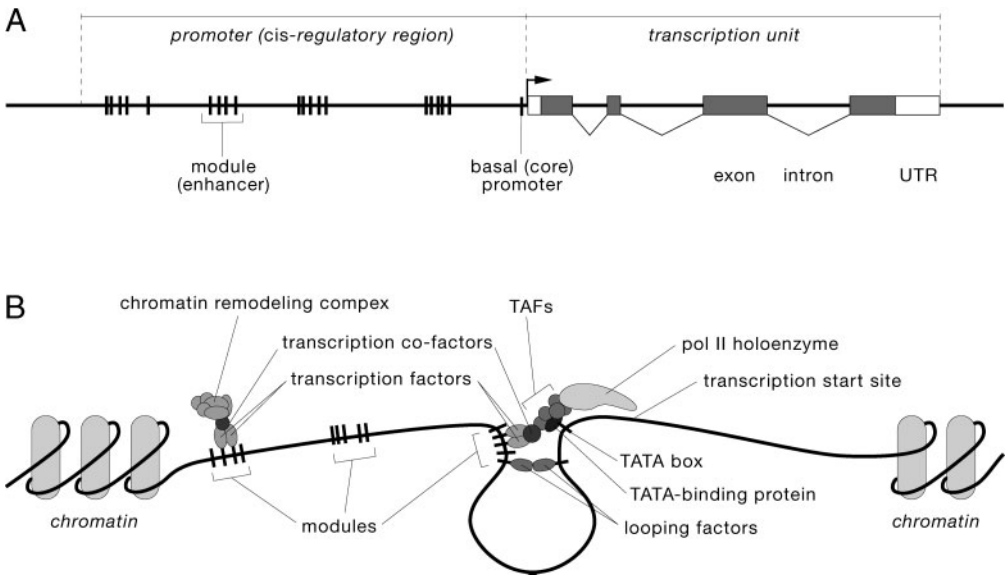

FIG. 1.—Promoter structure and function. (*A*) Organization of a generalized eukaryotic gene, showing the relative position of the transcription unit, basal promoter region (black box with bent arrow), and transcription factor binding sites (vertical bars). The position of transcription factor binding sites differs enormously between loci; although they often reside within a few kb 5′ of the start site of transcription (as shown here), many other configurations are possible (fig. 2). (*B*) Idealized promoter in operation. Initiating transcription requires several dozen different proteins which interact with each other in specific ways. These include the RNA polymerase II holoenzyme complex (~15 proteins); TATA-binding protein (TBP; 1 protein); TAFs (TBP-associated factors, also known as general transcription factors; ~8 proteins); transcription factors (precise composition and number bound differs among loci and varies in space and time and according to environmental conditions, but several to many any time transcription is active); transcription cofactors (again, precise composition and number will vary); and chromatin remodeling complexes (which can contain a dozen or more proteins).

a sequence-specific manner to the DNA near a gene, altering rates of transcriptional initiation. The shifting array of active transcription factors within the nucleus determines whether a gene is transcribed or not and how much mRNA is produced from it.

3.2 Promoter Structure

The organization of promoters is much less regular than that of coding sequences and lacks an equivalent of the genetic code or other sequence features that provide a consistent relationship to function. This fact has far-reaching implications for studying the evolution of promoter structure and function (see section 5).

3.2.1 Promoters Lack Universal Structural Features

No consistent sequence motifs exist for promoters of protein-coding genes. Two functional features are always present (fig. 1*A*), although they cannot always be recognized from sequence information alone. One is a basal promoter (or core promoter), the site upon which the enzymatic machinery of transcription assembles. Although necessary for transcription, the basal promoter is apparently not a common point of regulation, and it cannot by itself generate functionally significant levels of mRNA (Kuras and Struhl 1999; Lee and Young 2000; Lemon and Tjian 2000). The other functional feature is a collection of diverse transcription factor binding sites that confer specificity of transcription. Proteins bound to these sites produce a scalar response, the frequency with which new transcripts are initiated (Latchman 1998; Davidson 2001; Locker 2001).

3.2.2 The Transcriptional Machinery Assembles on the Basal Promoter

Eukaryotic genes that encode proteins are transcribed by the RNA polymerase II holoenzyme complex, which is composed of 10 to 12 proteins (Orphanides, Lagrange, and Reinberg 1998; Lee and Young 2000). This transcriptional machinery assembles on the basal promoter, a ~100-bp region whose functions are to provide a docking site for the transcription complex and to position the start of transcription relative to coding sequences (Reinberg et al. 1998; Lee and Young 2000; Pugh 2001). Basal promoter sequences differ among genes. For many genes, the critical binding site is a TATA box, usually located about 25–30 bp 5′ of the transcription start site. However, many genes lack a TATA box and instead contain an initiator element spanning the transcription start site. So-called null basal promoters exist that contain neither a TATA box nor an initiator element, and some basal promoters that contain one or the other also contain additional protein binding sites for general transcription factors (Carey and Smale 2000; Ohler and Niemann 2001). A gene may have more than one basal promoter, each of which initiates transcription at a distinct position (fig. 2*J* and *K*), and both TATA and TATA-less basal promoters can be associated with alternate start sites of the same gene (Goodyer et al. 2001). The functional consequences of differences in basal promoter structure are not well understood, although genes with TATA-less basal promoters may generally be

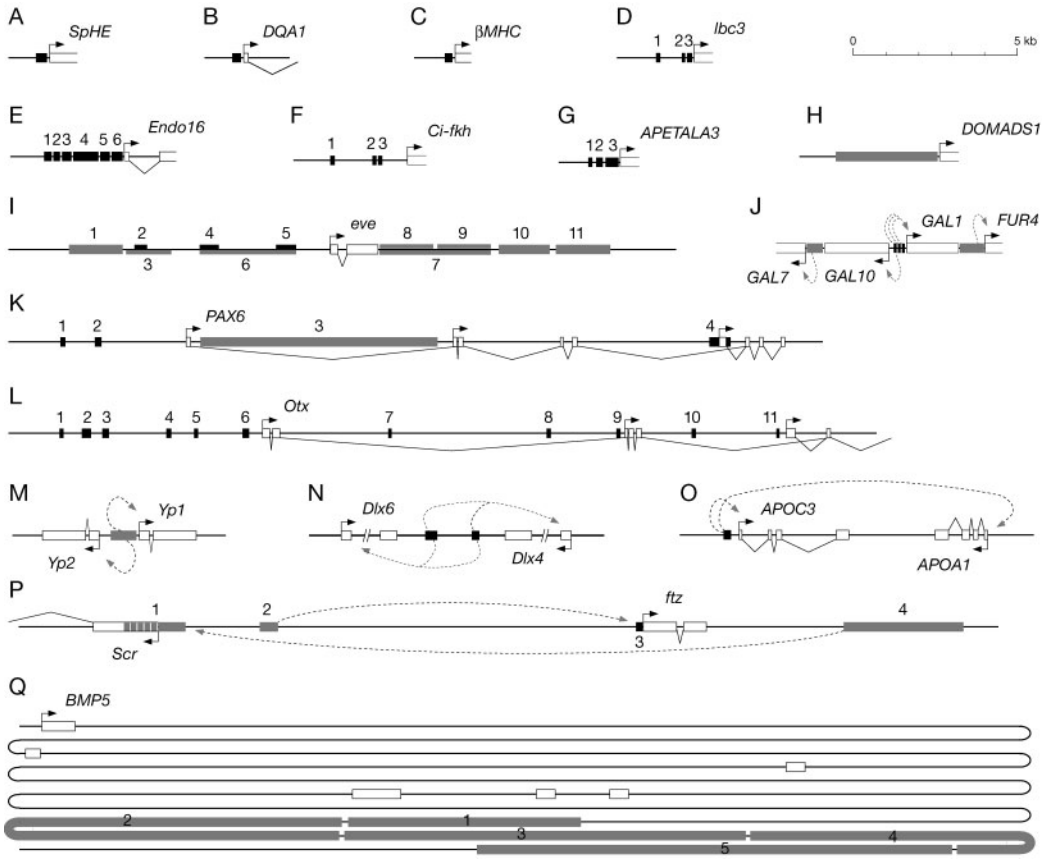

FIG. 2.—A bestiary of promoters. The known *cis*-regulatory regions of several genes from diverse eukaryotes are shown (partial promoters in panels *M* and *N*). These locus maps are drawn to the same scale (upper right): black boxes = precisely mapped regulatory regions; gray boxes = regions containing regulatory sequences that have not been mapped precisely (the actual extent of regulatory sequences is likely to be much smaller); white boxes = exons (UTRs and coding sequences); bent arrows = transcription start sites; numbers = distinct regulatory regions whose contribution to the total transcription profile has been defined experimentally; dashed lines indicate interactions between a module and more than one locus or a nonadjacent locus. Note the wide range in the spatial extent and position of *cis*-regulatory sequences. The smallest promoters of polymerase II–transcribed genes are in the range of 200–300 bp (*A*, *B*, *J*); in exceptional cases regulatory modules may lie more than 200 kb from the start of transcription (*Q*). Transcription factor binding sites generally reside in 5′ flanking sequences (*E–H*), but may also lie in the 5′ UTR (*P*: *Scr* module 1), introns (*L*: *Otx* modules 7–11), and 3′ flanking sequences (*Q*: *BMP5* modules 1–5). Nearly all promoters are compact in *Saccharomyces cerevisiae* (*J*), but promoter size differs enormously between loci in the sea urchin *Strongylocentrotus purpuratus* (compare *A*, *E*, *L*). Many promoters are highly modular, with different regulatory regions producing discrete components of the transcription profile (*D–F*, *I*, *K*). Some modules regulate transcription at more than one time and place during development (*G*, *K*: APETALA3 module 3 and *eve* modules 8–11). Conversely, some expression domains are regulated by more than one module (*I*: *eve* modules 3–7, 10, and 11 are required to produce the seven embryonic stripes of *eve* transcription; *K:* modules 1 and 4 are required for transcription of *PAX6* in the retina). Genes expressed in similar patterns sometimes have rather different promoter organization (*P*: module 2 of *ftz* produces seven embryonic stripes of transcription that are very similar to the ones produced by modules 3–7, 10, and 11 of *eve*). Although the *cis*-regulatory sequences of a given locus generally lie between it and the two flanking loci, in unusual cases there may be an intervening transcription unit. For instance, *ftz* lies between *cis*-regulatory sequences that interact only with *Scr* (*P*: module 4). In some cases, regulatory regions influence transcription at more than one locus. These may be divergently or convergently transcribed tandem paralogs (*M*, *N*: *Yp1*/*Yp2* and *Dlx6*/*Dlx4*, respectively) or even genealogically unrelated adjacent loci (*J*, *M*: *GAL10*/*GAL1* and *APOC3*/*APOA1*). Loci, protein product, taxon, and references: (*A*) *SpHE* (metalloendoprotease) of the sea urchin *Strongylocentrotus purpuratus* (Wei et al. 1995); (*B*) DQA1 (histocompatibility protein) of *Homo sapiens* (Petronzelli et al. 1995); (*C*) β*MHC* (myosin heavy chain) of *Rattus rattus* (Wright et al. 1999); (*D*) *lbc3* (leghemoglobin) of *Glycine max* (Stougaard et al. 1987; (*E*) *Endo16* (cell adhesion protein) of *S. purpuratus* (Yuh, Bolouri, and Davidson 1998, 2001); (*F*) *forkhead* (winged-helix transcription factor) of the urochordate *Ciona intestinalis* (Di Gregorio, Corbo, and Levine 2001); (*G*) APETALA3 (MADS-box transcription factor) of *Arabidopsis thaliana* (Hill et al. 1998); (*H*) *DOMADS1* (MADS-box transcription factor) of the orchid *Dendrobium* cv Madame Thong-In (Yu, Yang, and Goh 2002); (*I*) *even-skipped* (homeodomain transcription factor) of *D. melanogaster* (Sackerson, Fujioka, and Goto 1999); (*J*) *GAL10* and *GAL1* (genealogically unrelated metabolic enzymes) of *Saccharomyces cerevisiae* (West, Yocum, and Ptashne 1984); (*K*) *PAX6* (paired-box transcription factor) of *Mus musculus* (Kammandel et al. 1999); (*L*) *Otx* (homeodomain transcription factor) of *S. purpuratus* (Yuh et al. 2002); (*M*) *Yp1* and *Yp2* (paralogous yolk proteins) of *D. melanogaster* (Chung et al. 1996); (*N*) *Dlx6* and *Dlx4* (paralogous homeodomain transcription factors) of *Danio rerio* (Zerucha et al. 2000; the intron/exon structure of the loci is not known in detail; only shared regulatory elements are shown); (*O*) *APOC3* and *APOA1* (genealogically unrelated lipid carrier proteins; only shared regulatory elements are shown) of *H. sapiens* (Li et al. 1995; Naganawa et al. 1997); (*P*) *ftz* and *Scr* (paralogous homeodomain transcription factors) of *D. melanogaster* (Calhoun, Stathopoulos, and Levine 2002); (*Q*) *BMP5* (signaling protein) of *M. musculus* (DiLeone, Russell, and Kingsley 1998; the position of exon 3 is not known precisely; splice patterns are omitted for simplicity).

transcribed constitutively at relatively low levels (Pugh 2001). A key early step in transcriptional initiation is attachment of TATA-binding protein (TBP) to DNA (Jackson-Fisher et al. 1999; Kuras and Struhl 1999). In promoters lacking TATA boxes, proteins that associate with other basal promoter motifs facilitate TBP association with DNA in a sequence-independent manner. Once TBP binds, several TBP-associated factors (TAFs) guide the RNA polymerase II holoenzyme complex onto the DNA (fig. 1*B*). This step, which can be positively or negatively modulated by transcription factors bound at other sites, is one of the most important points of transcriptional regulation (Latchman 1998; Lee and Young 2000; Lemon and Tjian 2000).

3.2.3 The Start Site of Transcription Varies in Both Sequence and Position

The start site of transcription, unlike the start site of translation, does not require a specific sequence motif and cannot be identified from sequence data. After the RNA polymerase II holoenzyme complex assembles onto DNA, a second contact is established ~30 bp downstream. This second contact point is the start site of transcription. It is thus the physical size of the transcriptional machinery and the particular composition of binding sites that facilitate its binding to the basal promoter and that determine where transcription begins (fig. 1*B*). Spacing between the start sites of transcription and translation differs considerably among genes, ranging from $\sim 10^1$ to 10^4 bp; the 5′ untranslated region (UTR) can also contain introns that alter its length post-transcriptionally. The functional consequences of differences in 5′ UTR length are not well understood.

3.2.4 Basal Promoters Provide Limited Transcriptional Activity and Specificity

By itself, a basal promoter initiates transcription at a very low rate, even when the local chromatin is suitably decondensed (Jackson-Fisher et al. 1999; Kuras and Struhl 1999; Lemon and Tjian 2000). Furthermore, most of the proteins that bind to basal promoter motifs are ubiquitously expressed and therefore provide little regulatory specificity (Carey and Smale 2000; Lee and Young 2000; Lemon and Tjian 2000). These proteins are known as general transcription factors. A few tissue-specific isoforms of these proteins are known, however, and may exert some degree of transcriptional regulation (Holstege et al. 1998; Smale et al. 1998). Additional mechanisms of transcriptional regulation involving the basal promoter are discussed later (see section 3.3.6).

3.2.5 Specificity of Transcription Is Controlled by Proteins that Bind to Discrete, Idiosyncratic Sites

Producing functionally significant levels of mRNA requires the sequence-specific association of transcription factors with DNA sequences outside the basal promoter (Weinzierl 1999; Carey and Smale 2000; Lemon and Tjian 2000). The composition and organization of these transcription factor binding sites varies enormously among genes (fig. 2). The nucleotide sequences of these binding sites determine which transcription factors are capable of associating with the promoter of a given gene. Which transcription factors actually do so depends on which of them is present in the nucleus in an active form and, in many cases, on the presence of cofactors as well (Locker 2001). The complement of active transcription factors within the nucleus differs during the course of development, in response to environmental conditions, across regions of the organism, and among cell types (Latchman 1998; Davidson 2001). This changing array of transcription factors provides nearly all of the control over when, where, at what level, and under what circumstances a particular gene is transcribed. Thus, the genetic basis for the expression profile of each gene resides in part within its promoter and in part within the many other segments of the genome that encode specific transcription factors that bind to the promoter.

3.3 Transcription Factor Binding Sites

The composition and configuration of transcription factor binding sites near a gene are major determinants of its expression profile, and they therefore constitute an important class of sequences that are potential targets of natural selection on gene expression.

3.3.1 Promoters Contain Numerous Transcription Factor Binding Sites

Identifying genuine binding sites is not straightforward for a variety of reasons (see sections 3.3.3 and 5.2; Weinzierl 1999; Carey and Smale 2000). It is difficult to be certain that all functional binding sites within a promoter have been identified, and it is prudent to assume that some binding sites remain uncharacterized even within well-studied promoters. Because of this uncertainty, the range and average number of binding sites found in a typical promoter is not known, much less any correlations between these parameters and the nature of the gene product or mode of expression. Nonetheless, a perusal of well-characterized eukaryotic promoters suggests that numbers on the order of 10–50 binding sites for 5–15 different transcription factors is not unusual (for examples, see Arnone and Davidson [1997] and Wilkins [2002]).

3.3.2 Transcription Factor Binding Sites Are Distributed Sparsely and Unevenly

Binding sites typically comprise a minority of the nucleotides within a promoter region. This fraction ranges from 10% to 20% within relatively well-studied regulatory regions (table 1, fig. 3). These regions are often interspersed with regions that contain no binding sites (fig. 2). Disjunct regulatory regions often produce discrete portions of the total transcription profile (see section 3.5.4). Nucleotides that do not affect the specificity of transcription factor binding are generally assumed to be nonfunctional with respect to transcription. In some cases, however, these nucleotides may influence the local conformation of DNA, with direct consequences for protein binding (e.g., Naylor and Clark 1990; Hizver et al. 2001; Rothenburg et al. 2001). Spacing between

Table 1
Density of Binding Site Nucleotides in Promoter Regions

Locus/Species	Region[a]	Binding/Nonbinding[b]	Proportion Binding	Reference[c]
Endo16, Strongylocentrotus purpuratus	Module A	33/130	0.22	1
leghemoglobin, Glycine max		23/191	0.11	2
Adh, Arabidopsis thaliana		60/440	0.12	3
Adh, Drosophila melanogaster		345/1155	0.23	4
even-skipped, Drosophila melanogaster	Stripe 2 module	285/1430	0.10	5

[a] Few promoters have been analyzed in sufficient detail that nucleotides over their entire extent can be confidently assigned to binding sites versus nonbinding sites (see section 5.2). For all the examples shown here, the promoter is larger (in some cases much larger) than the region for which detailed information is available (fig. 2).

[b] Nucleotides identified by the authors as involved in specific binding of transcription factors, as a fraction of all nucleotides comprising the module or promoter region. Somewhat different criteria were used to identify binding sites in these studies, and tallies of binding site nucleotides are likely to be underestimates (see section 5.2).

[c] References: (1) Yuh, Bolouri, and Davidson 2001; (2) Stougaard et al. 1987, Andersson et al. 1996; (3) Miyashita 2001; (4) Nurminsky et al. 1996; (5) Small, Blair, and Levine 1992.

binding sites varies enormously, from partial overlap to tens of kilobases (figs. 2 and 3). Functional constraints on binding site spacing are often related to protein interactions that take place during DNA binding (see section 3.5.2).

3.3.3 Transcription Factor Binding Sites Are Short and Imprecise

Because of the way transcription factors interact with DNA, several different criteria are used to define binding sites. (1) *Physical contact versus binding specificity*. The segment of DNA protected from nuclease digestion by a transcription factor (its ''footprint'') is typically wider than the nucleotides that confer binding specificity (its binding site). Most transcription factor binding sites span 5–8 bp (table 2), whereas footprints are typically 10–20 bp. (2) *Single versus multiple sequences*. Most binding sites can tolerate at least one, and often more, specific nucleotide substitution without completely losing functionality (Latchman 1998; Courey 2001). This is evident from comparing different binding sites known to bind the same transcription factor and from in vitro assays of protein-DNA binding (see section 3.4.4; for examples within a single promoter, see fig. 3). The full range of sequences (in practice, often poorly understood) that can bind a particular transcription factor with significantly higher specificity than random DNA under physiological conditions is often described by a position weight matrix, in which the probability that each position in the binding site will be represented by a particular nucleotide is tabulated. When binding site matrices are factored in, the number of nucleotides required for specific protein binding drops to about 4–6 bp for a typical binding site (table 2). Although binding site matrices are generally composed of related sequences, some transcription factors bind to rather different sequences in association with different binding partners (e.g., jun/jun, fos/jun, CRE-BP1/jun dimers: Latchman 1998, Fairall and Schwabe 2001). (3) *Informatic versus functional consensus*. The term *consensus sequence* refers to the single ''best'' variant of the binding site matrix or to a degenerate sequence that captures most of the binding site matrix (table 2). Two rather different criteria are used to define consensus sequences: sequence comparisons (most commonly, simply the average sequence of multiple instances of binding sites for same protein) and biochemical assays (the single variant with the highest affinity for the protein in vitro).

3.3.4 Many Potential Binding Sites Are Nonfunctional

Given that there are many different transcription factors with different binding matrices, and given that binding sites are short and imprecise, every kilobase of genomic DNA contains many dozens of potential transcription factor binding sites on the basis of random similarity (Carroll, Grenier, and Weatherbee 2001; Stone and Wray 2001). For a variety of reasons (fig. 4), many of these consensus matches don't bind protein in vivo and have no influence on transcription (Biggin and McGinnis 1997; Weinzierl 1999; Li and Johnston 2001). Identifying the potential binding sites that actually bind protein requires biochemical and experimental tests (see sections 5.2 and 5.3).

3.3.5 Variants Within a Binding Site Matrix Can Differ Functionally

Although most transcription factors can bind to several distinct sequences, they may do so with different kinetics (Czerny, Schaffner, and Busslinger 1993; Carey and Smale 2000). Differences in binding affinities are particularly important when two binding sites overlap physically or are located very near each other, because only one binding site can be occupied by protein at a time (fig. 3*A*: Otx, Z, and CG binding sites). In such cases, differences in protein concentrations and binding kinetics will determine which binding site is occupied most of the time. Differences in kinetics can also be important for binding sites not near each other, because active promoters compete for a single pool of transcription factors within each nucleus and there are typically fewer transcription factors present than there are binding sites in a genome.

3.3.6 Transcription Factor Binding Sites Occupy a Wide Range of Positions Relative to the Transcription Unit

Although transcription factor binding sites sometimes occupy a single, discrete region near the start site of transcription (fig. 2*A–E*), in many cases they are dispersed

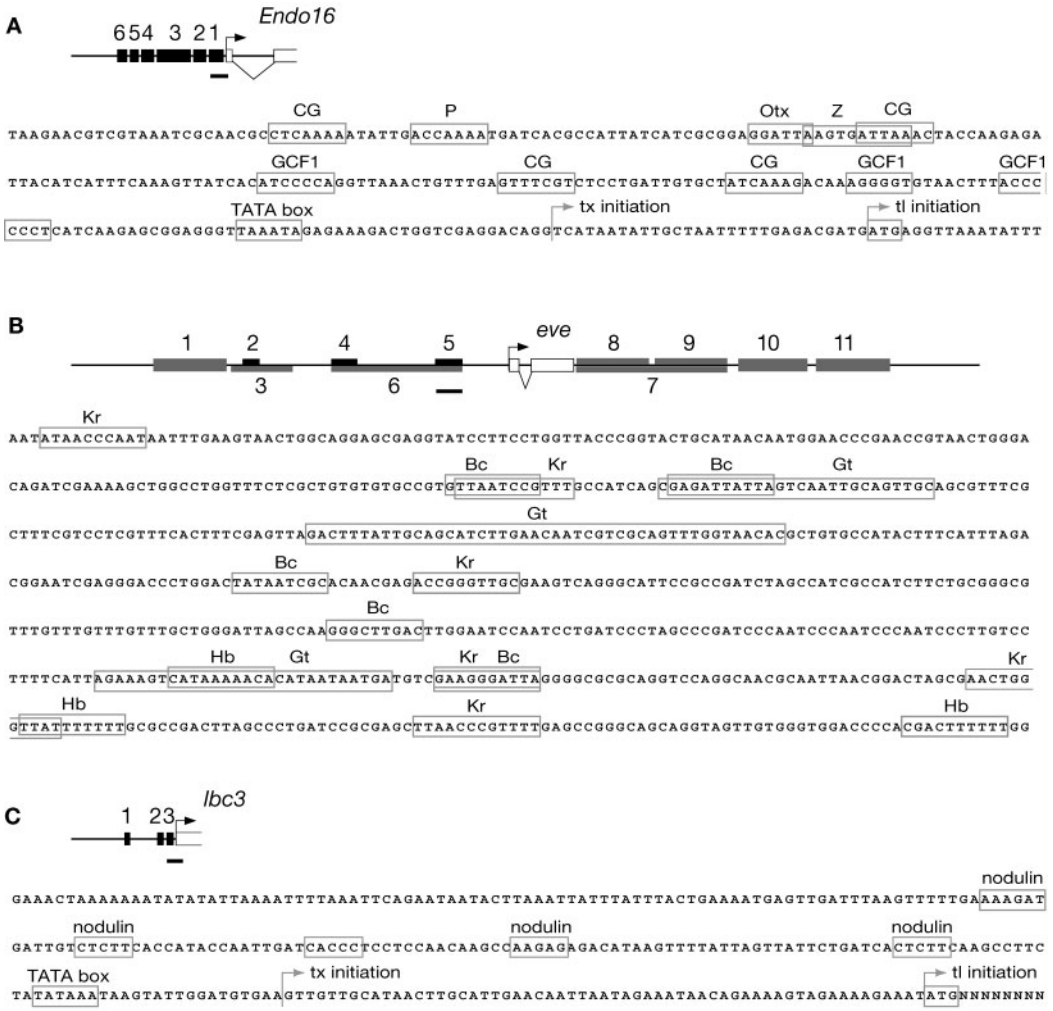

FIG. 3.—Examples of binding site organization. Transcription factor binding sites within three *cis*-regulatory regions are shown to scale. Boxes indicate nucleotides that contribute to binding specificity (with the exception of Gt, where footprints are shown); transcription factor names or binding site motifs are shown above binding sites; tx initiation = start site of transcription; tl initiation = start site of translation; solid bars under inset maps of locus organization indicate the approximate position of the sequence shown. Note that nucleotides contributing to protein binding comprise only a small fraction of the total, even within those regions where binding sites are relatively dense (see also table 1). Multiply-represented binding sites are often present in both orientations and represent variations of the consensus (CG and GCF1 of *Endo16*; Kr and Bc of *eve*; CTCTT nodulin motif of *lbc3*). Spacing between binding sites provides clues to function: those centered ~10 bp apart may bind proteins that interact on the same side of the DNA helix (Otx and CG of *Endo16*; upstream nodulin motifs of *lbc3*); those that overlap may operate as a switch, with one protein preventing the other from binding under some conditions (Otx, Z, and CG of *Endo16*; Kr and Bc of *eve*); whereas those more than about 20 bp apart probably bind proteins that either do not interact or do so through DNA bending or looping. (*A*) Module A and basal promoter of *Endo16* from the sea urchin *Strongylocentrotus purpuratus* (Yuh, Bolouri, and Davidson 1998). Module A contains 11 binding sites for five transcription factors, two of which interact with multiple binding sites. (*B*) Stripe 2 module from *eve* of *D. melanogaster* (Small et al. 1992). This module contains 17 binding sites for four transcription factors, all of which interact with multiple binding sites. Note that boxed nucleotides for Gt are footprints rather than binding sites. (*C*) Proximal promoter region of *lbc3* of *Glycine max* (Stougaard et al. 1987). This region contains five binding sites for at least three different transcription factors. Multiple instances of two different nodulin motifs are present; these binding sites are found upstream of several genes that are expressed in root nodules of legumes.

into several distinct clusters (fig. 2*I*, *K–L*, and *P*). The physical extent of *cis*-regulatory regions varies by nearly three orders of magnitude, from a few hundred base pairs to >100 kb (fig. 2). An extreme example of physical dispersion is a regulatory module of the *Shh* locus in humans and mice that lies ~800 kb distant from the start site of transcription (Lettice et al. 2002). The position of transcription factor binding sites relative to the transcription unit also differs enormously among genes. They often lie within a few kb 5′ of the basal promoter (fig. 2*A–G*), but they can occupy a wide range of other positions: > 30 kb 5′ of the basal promoter (e.g., *Ubx* in *D. melanogaster*: Simon et al. 1990; *Pax-6* in mouse: Kammandel et al. 1999; *APOB* in humans: Nielsen et al. 1998); within the 5′ UTR (*Scr* in *D. melanogaster*: Calhoun, Stathopoulos, and Levine 2002); within introns (*Otx* in the sea urchin *Strongylocentrotus purpuratus*: Yuh et al. 2002; *CCR5* in humans: Bamshad et al. 2002); > 30 kb 3′ of the transcription unit (*BMP5* in mouse: DiLeone, Russell, and Kingsley 1998); and, in rare instances, even within a coding exon (*keratin 18* in humans: Neznanov, Umezawa, and Oshima 1997; *nonA* in Drosophila: Sandrelli et al. 2001). This diversity of positions is possible because DNA looping allows interaction between

Table 2
Size and Information Content of Transcription Factor Binding Sites

Transcription Factor	Consensus Binding Site[a]	Information Content[b]	Reference[c]
C/EBP	RTTGCGYAAY	17 bits	1
Runt	TGYGGTY	12 bits	2
Krox-20	GCGGGGCG	16 bits	3
Otx	RGATTA	11 bits	4
eve	ATTA	8 bits	5
Pax-5	RNNCANTGNNGCGKRACSR	23 bits	6
AP1	CCWWWWWWGG	14 bits	7
Myc-bHLH	CACGTG	12 bits	8
TATA-binding protein	TATAWAW	12 bits	3
MNF1	CCRCCC	11 bits	9
Jun/Fos heterodimer	TGAGTCA	14 bits	10
Jun/CREB heterodimer	TGACGTCA	16 bits	10
PBC/Hox heterodimer	TGATNNATTA	16 bits	11

[a] Consensus sequence, as reported by the authors of the reference in the right-hand column. R = G/A, W = A/T, Y = C/T, K = G/T, S = C/G, N = A/C/G/T.

[b] Each nonredundant nucleotide position contains two bits of information (i.e., can be represented by two binary states); similarly, twofold redundant positions contain one bit.

[c] (1) Osada et al. (1996); (2) Kramer et al. (1999); (3) Latchman (1998); (4) Klein and Li (1999); (5) Biggin and McGinnis (1997); (6) Czerny, Schaffner, and Busslinger (1993); (7) Riechmann, Wang, and Meyerowitz (1996); (8) Coller et al. (2000); (9) Morishima (1998); (10) Benbrook and Jones (1990); (11) Lufkin (2001).

proteins associated with DNA at distant binding sites (fig. 1*B*) (see section 3.4.5). Binding sites may even lie on the far side of an adjacent locus (fig. 2*O*). The position of binding sites for some transcription factors may be functionally constrained. For instance, CCAAT binding sites for the transcription factor CBP (CREB binding protein) are generally located 50–100 bp 5′ of the transcription start site, and those for Sp1 are often located near the basal promoter of many mammalian genes. For most transcription factors, however, binding sites lack any obvious spatial restriction relative to other features of the locus. In general, the functional consequences of binding site position are poorly understood.

3.3.7 Specific Sequences Limit the Regulatory Influence of Binding Sites

Because binding sites can interact with basal promoters that are tens or even hundreds of kilobases distant, they are potentially able to influence transcription at more than one locus. At least three mechanisms spatially restrict this influence. (1) *Insulator sequences*. Some, and perhaps many, promoters are bounded by insulator sequences (also known as boundary elements) (Wolffe 1994; Bell and Felsenfeld 1999; Dillon and Sabbattini 2000). Mechanisms of insulator function are not well understood but appear to involve chromatin modulation (Bell and Felsenfeld 1999). (2) *Basal promoter selectivity*. Some regulatory sequences interact preferentially with TATA or TATA-less basal promoters, even if a basal promoter of the other kind is closer to them (Ohtsuki, Levine, and Cai 1998). (3) *Selective tethering*. Sequences immediately 5′ of a basal promoter may help selectively recruit transcription factor complexes bound at distant sites. For instance, an activator module (enhancer) located close to the *ftz* locus in Drosophila associates only with the more distant basal promoter of *Scr* (fig. 2*O*) (Calhoun, Stathopoulos, and Levine 2002).

3.3.8 Some Binding Sites Affect Transcription at More than One Locus

Although most binding sites directly influence the expression of just one gene, many exceptions are known. One manifestation is a ''divergent promoter,'' where binding sites regulate transcription of paralogous loci that lie on opposite strands of DNA with their 5′ ends centrally located (fig. 2*M*). Binding site ''sharing,'' or cross-regulation, of adjacent loci also occurs in other contexts: paralogs that are transcribed convergently (fig. 2*N*) or in parallel (e.g., beta-globin: Grosveld et al. 1993; *Hox* complex: Ohtsuki, Levine, and Cai 1998; Kmita, Kondo, and Duboule 2000) and even among genealogically unrelated loci that lie near each other (fig. 2*J* and *O*). Single mutations in single binding sites may affect the transcription of more than one gene. In humans, for example, segregating variants are known that simultaneously influence transcription of the genes encoding beta-globin and gamma-globin (Metherall, Gillespie, and Forget 1988; Grosveld et al. 1993), the insulin and *IGF2* genes (Paquette et al. 1998), and the *APOA1* and *APOCIII* genes (Li et al. 1995; Naganawa et al. 1997). In the last case (fig. 2*O*), nucleotide variants have distinct effects on each locus: the rare haplotype downregulates *APOA1* in the colon but upregulates *APOCIII* in the liver. Cross-regulation may be the reason for the long-term physical linkage of genes in the *Hox* complexes of animals (Lufkin 2001). The general prevalence of cross-regulation remains uncertain (Bonifer 2000). Even where cross-regulation is known to occur, however, the involved loci are sometimes each regulated by unique regulatory sequences as well as shared ones, providing some degree of differential regulation.

3.4 Transcription Factors

The transcription of every gene is regulated by transcription factors and cofactors that interact with its promoter. The distant and dispersed regions of the genome

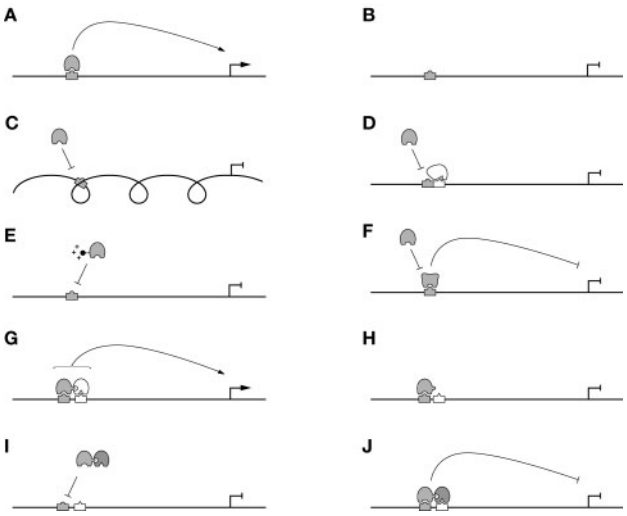

FIG. 4.—Context-dependence of transcriptional regulation. The function of a transcription factor binding site is always context dependent to some extent. (*A*) The binding site for a protein that activates transcription, for instance, will not function under several different conditions: (*B*) when the transcription factor is absent; (*C*) when local chromatin is condensed, whether or not the transcription factor is present; (*D*) when an adjacent binding site is occupied, masking the binding site of interest; (*E*) when the transcription factor is present but in an inactive form; or (*F*) when a different protein is present that has a higher affinity for the binding site. (*G*) Many transcription factors interact with cofactors to exert their influence on transcription. In such cases, additional situations contribute to context dependence and the binding site will not function or will function differently: (*H*) when the cofactor is absent; (*I*) when a different cofactor is present that alters binding specificity; or (*J*) when a different cofactor is present that allows binding but alters subsequent protein interactions.

that encode these proteins constitute a second important class of sequences that are potentially the target of natural selection on the transcription profile of a particular gene.

3.4.1 Transcription Factors Belong to a Relatively Small Number of Gene Families

Most transcription factors belong to gene families (Latchman 1998; Locker 2001). The size of each transcription factor gene family differs considerably among genomes (table 3), but the reasons and functional consequences of these differences are not understood. Existing paralogs are the result of duplications that occurred across a wide range of times, from before the divergence of eukaryotic kingdoms to much more recently (Duboule 1994; Bharathan et al. 1997; Dailey and Basilico 2001; Stauber, Prell, and Schmidt-Ott 2002). There are approximately 12 to 15 structurally distinct DNA-binding domains known from eukaryotic transcription factors (Harrison 1991; Fairall and Schwabe 2001). For intensively studied organisms, the known transcription factor families may constitute a nearly complete list. Far less is known about the diversity and evolutionary history of transcription cofactors, proteins that bind to transcription factors but not to DNA (fig. 1*B*; see the following section and section 3.4.5).

3.4.2 Transcription Factors Are Structurally and Functionally Modular Proteins

Most transcription factors contain several distinct functional domains. These may include almost any combination of the following. (1) *DNA-binding domains.* The amino acids that comprise the DNA binding region may be contiguous (e.g., homeodomain, MADS box) or dispersed within the primary sequence (e.g., Zn-fingers). Some transcription factors contain two distinct DNA binding regions (e.g., many *Pax* family members contain both a homeodomain and a paired-box domain). (2) *Protein-protein interaction domains.* Transcription factors engage in a variety of interactions with other proteins (see section 3.4.5). Most transcription factors contain from one to several such domains. Interaction domains, which generally are more difficult to recognize from sequence inspection than DNA binding domains, include leucine zippers and the pentapeptide motif of homeodomain proteins (Latchman 1998). (3) *Domains that act as intracellular trafficking signals.* Many transcription factors contain a nuclear localization signal. In some cases, the activity of a transcription factor may be modulated by controlling the ratio of cytoplasmic-to-nuclear localization (e.g., Exd: Abu-Shaar, Ryoo, and Mann 1999). (4) *A ligand-binding domain.* Some transcription factors, such as specific steroid hormones, can bind ligands which modulate their activity. Most known cases belong to the nuclear receptor family (Benecke, Gaudon, and Gronemeyer 2001), but an unrelated Ca^{2+}-binding transcription factor has recently been discovered (Carrión et al. 1999).

Many protein-DNA binding domains predate the divergence of plants and animals (e.g., homeodomain: Bharathan et al. 1997), as do some protein-protein interaction domains (Bürglin 1997). The evolutionary history of transcription factor gene families includes many examples of ''domain shuffling'' and loss of specific domains. For instance, a paralog may retain a DNA-binding domain but lose a protein-protein interaction domain responsible for transcriptional activation; the resulting protein will function as a repressor if it competes for binding sites with a paralog that contains an activation domain (e.g., Sp family: Suske 1999). Transcription cofactors, by definition, lack a DNA-binding domain, but they typically contain domains that mediate a specific protein-protein association with a transcription factor and directly or indirectly interact with effector complexes (either the transcriptional machinery or chromatin remodeling complexes).

3.4.3 Transcription Factor Structure Determines DNA Binding Specificity

The DNA binding domain of most transcription factors is a short motif, most commonly an alpha helix but sometimes a beta-strand or a less organized loop, that inserts into the major groove of double-stranded DNA (Choo and Klug 1997; Jones et al. 1999; Fairall and Schwabe 2001). A single amino acid substitution within the binding domain can alter binding specificity (Treisman et al. 1989; Mathias et al. 2001). DNA binding domains are often highly conserved evolutionarily (Duboule 1994;

Table 3
Size of Selected Transcription Factor Families in Five Eukaryotes

Transcription[a] Factor Family	*Saccharomyces cerevisiae*	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Homo sapiens*	*Arabidopsis thaliana*
Homeodomain	9	109	148	267	118
Nuclear receptor	1	183	25	59	4
Zn-finger	121	437	357	706	1,049
Runt-domain	0	2	4	3	0
Basic HLH	7	41	84	131	106
Paired box	0	23	28	38	2
Myb	15	17	18	32	243

[a] Tallies of number of genes in each family from Venter et al. (2001) and Lander et al. (2001).

Dailey and Basilico 2001), although functional polymorphisms that lead to differences in binding kinetics are known (e.g., Brickman et al. 2001). Sequence-specific protein-DNA contacts rarely extend across more than 5 bp, and for some motifs, such as Zn-fingers, they extend only 3 bp. The extent of this physical interaction is not sufficient to provide much sequence specificity, as a given 5-bp sequence occurs on average every 1,024 bp. Three structural features can increase DNA binding specificity (Latchman 1998; Fairall and Schwabe 2001): (1) *multiple DNA binding domains can exist within a single transcription factor* (e.g., most Pax family members contain both paired-box and homeodomain DNA binding domains, whereas all Zn-finger transcription factors contain multiple Zn-fingers); (2) *additional structural features can bind nearby nucleotides through minor groove contacts* (e.g., many homeodomain and GATA factors); and (3) *binding to DNA may require homodimerization or heterodimerization* (e.g., myc/mad/max, fos/jun, and most nuclear receptor family members). All three structural features effectively increase the number of specific nucleotides required for efficient binding and typically involve noncontiguous nucleotides within promoters (table 2).

3.4.4 Transcription Factors Bind to More than One Sequence, Although They Do So with Different Affinities

Transcription factors bind relatively tightly to double-stranded DNA (K_d is typically in the range of 10^{-9} to 10^{-10}), with a high degree of sequence specificity (Biggin and McGinnis 1997; Carey and Smale 2000). Because of their sequence specificity and binding kinetics, and because many potential target sites are present in a genome, eukaryotic transcription factors need to be present in copy numbers of $\sim$5–20 $\times$ 10^3 per nucleus in order to bind efficiently (Dröge and Müller-Hill 2001). Although they associate in a sequence-specific manner, most transcription factors bind a range of motifs rather than a single one (see section 3.3.3). The extent of this binding site matrix differs considerably among transcription factors (table 2). Binding specificity may be strongly influenced by cofactors. For instance, some Hox transcription factors interact with TALE family proteins, resulting in more efficient binding or in binding to a narrower consensus (Knoepfler and Kamps 1995; Berthelsen et al. 1998). Post-translational modifications, most commonly phosphorylation, can also modulate binding specificity. Several enzymes, including the MAP and Janus kinases, fine-tune the phosphorylation state of transcription factors, exerting a significant influence on overall transcription patterns (Shore and Sharrocks 2001). Paralogous transcription factors may interact with the same binding site (table 4), although their binding kinetics may differ. The consensus sequence for most transcription factors is not yet well defined, with most consensus determinations based on sequence comparisons rather than direct biochemical or functional analyses. Surprisingly little information exists about evolutionary changes in consensus sequences.

3.4.5 Transcription Factors Influence Transcription Through Protein-Protein Interactions

All proteins that regulate transcription directly or indirectly influence the frequency with which the polymerase II complex assembles onto the basal promoter. This influence is exerted through a wide variety of protein-protein interactions, the most common of which are summarized in figure 1 and discussed below (Latchman 1998; Courey 2001; Shore and Sharrocks 2001). In general, the protein-protein interaction domains of transcription factors are not as well characterized as their DNA-binding domains. (1) *A transcription factor bound to DNA can interact with components of the basal transcriptional machinery, facilitating or inhibiting its association with the basal promoter and resulting in an increase or decrease in overall transcription rates*. These interactions are specific and take place through protein-protein interaction domains (Triezenberg 1995; Torchia, Glass, and Rosenfeld 1998). Some transcription factors contain activation domains that associate directly with one of the TAFs (TBP-associated factors, also known as general transcription factors) to increase the frequency with which the RNA polymerase II complex initiates transcription (e.g., GAL4: Gill and Ptashne 1987), whereas others contain repression domains that have the opposite effect (e.g., engrailed: Jaynes and O'Farrell 1991). (2) *A transcription factor may interact with another transcription factor before or as it binds to DNA, in a variety of functional contexts.* Several transcription factors bind DNA only as homodimers or heterodimers (e.g., many nuclear receptor family members: Benecke, Gaudon, and Gronemeyer 2001); others can bind DNA only when they are not bound to a cofactor (e.g., myoD and Id: Benezra et al. 1990); and still others can bind DNA alone, but their specificity and/or association kinetics change when

Table 4
Overlapping Binding Site Specificities

Binding Site	Transcription Factors that Bind[a]	Reference[b]
Paralogs[c]		
ATTA	Engrailed, even-skipped, fushi-tarazu	1
GGATTA	Orthodenticle, goosecoid	2
CACGTG	Myc, Mad	3
Unrelated Proteins[d]		
CCATATTTGG	SRE, YY1[e]	4
TCAATGT	IRE-ABP, C/EBP[e]	5
GGGGCGTGGGCTG	Sp1, Egr[e]	4

[a] Even though these proteins all bind specifically to the sequences shown, binding kinetics may differ. Proteins are probably recognizing a subset of nucleotides in the longer target sites, based on their known specificities for other sequences.

[b] (1) Biggins and McGinnis (1997); (2) Angerer et al. (2001); (3) James and Eisenman (2002); (4) Fry and Farnham (1999); (5) Buggs et al. (1998).

[c] Proteins belonging to the same family; not necessarily the closest paralogs.

[d] Proteins with no discernible genealogical relationship.

[e] Overline is binding site of first protein listed, underline of second protein.

complexed with a cofactor (e.g., many homeodomain proteins: Lufkin 2001). Other transcription factors bind as heterodimers with a variety of partners, with distinct consequences for transcription (e.g., homeodomain proteins: Pinsonneault et al. 1997; max and partners myc/mad: Grandori et al. 2000). (3) *A transcription factor bound to DNA may physically inhibit binding of a different transcription factor to a nearby site*. For steric hindrance to work, the two binding sites must be near each other (usually on the same face of the DNA strand), and the affinity of the blocking protein for its binding site or its concentration must exceed that of the blocked protein. Because steric hindrance involves nonspecific protein-protein interactions, in principle any transcription factor can operate in this way. (4) *A transcription factor bound to DNA may alter chromatin structure*. Some transcription factors maintain local chromatin in a decondensed state (e.g., trithorax: Mahmoudi and Verrijzer 2001), and others condense it (e.g., groucho: Chen and Courey 2000; polycomb: Jacobs and van Lohuizen 1999). These transcription factors recruit multiprotein complexes such as the SWI/SNF complex (Varga-Weisz 2001), enzymes that acetylate, deacetylate, methylate, or demethylate histones (Vogelauer et al. 2000; Richards and Elgin 2002), or enzymes that methylate or demethylate DNA (Jones and Takai 2001). Chromatin remodeling is highly dynamic and is apparently regulated on spatial scales as small as promoters, or even regions within a promoter (Kadosh and Struhl 1998; Wolffe 2001). Although chromatin condensation probably overrides most protein-DNA interactions by physically blocking access to binding sites, some transcription factors can associate with DNA in partially condensed chromatin (Narlikar, Fan, and Kingston 2002). (5) *A transcription factor bound to DNA may stabilize the bending or looping of DNA*. Some proteins facilitate local bending of DNA, allowing other bound proteins that are near each other but not in contact to interact (e.g., Sox: Scaffidi and Bianchi 2001; Tcf/Lef-1: Love et al. 1995; Sp1: Sjottem, Andersen, and Johansen 1997). Other proteins stabilize DNA loops by forming homodimers, facilitating interactions among transcription factors bound at distant sites (e.g., GCF1: Zeller et al. 1995; RIP60: Houchens et al. 2000). Some of these so-called architectural proteins may be necessary, rather than sufficient, to activate or repress transcription; others, however, play a direct role in modulating the frequency of transcriptional initiation (e.g., Tcf/Lef-1: Fry and Farnham 1999). (6) *Transcription cofactors that do not bind to DNA can mediate interactions between DNA-binding proteins and the transcriptional machinery or chromatin-remodeling enzymes*. Some proteins influence transcription by mediating specific interactions between transcription factors and effector proteins, primarily cofactors of the polymerase II complex or chromatin-remodeling complexes. Many such transcription cofactors have been identified. Some seem to interact with a restricted set of transcription factors (e.g., OCA-B: Gstaiger et al. 1995; TIF-1: Glass, Rose, and Rosenfeld 1997), but others interact with a variety of phylogenetically unrelated transcription factors (e.g., CBP/p300 interacts with CREB, myoD, Myb, Jun, Fos, nuclear receptor, and AP-1 family members: Shikama, Lyon, and La Thangue 1997, Wolffe 2001). Promoter sequences contain no direct evidence of cofactor interactions.

3.4.6 Many Transcription Factors Act Primarily as Activators or Repressors of Transcription

The presence of particular protein-protein interaction domains dictates to a large extent what effect a given transcription factor will have once it is bound to DNA (see section 3.4.5). A variety of transcriptional activation domains have been identified that mediate direct interaction with TBP or indirect interaction, by means of a TAF (Triezenberg 1995). Some transcription factors contain more than one activation domain (e.g., GAL4: Gill and Ptashne 1987; CREB: White 2001). Likewise, various repressor domains are known, although their mechanisms of operation are less well understood (Hanna-Rose and Hansen 1996; Latchman 1998). ''Domain-swapping'' experiments demonstrate that these domains alone are sufficient to turn a transcription factor from an activator into a repressor and vice-versa.

3.4.7 The Effect of Some Transcription Factors Is Context Dependent

The activity of many transcription factors depends on post-translational covalent modifications, most commonly phosphorylation (e.g., Oct-1: Segil, Roberts, and Heitz 1991), acetylation (e.g., p53: Gu and Roeder 1997), and glycosylation (Sp1: Jackson and Tjian 1988). These modifications often provide an important point of control over transcription, and phosphorylation in particular is often dynamically regulated (Roberts, Segil, and Heintz 1991). The effect of a transcription factor may be strongly context dependent, even once it is bound to DNA and despite the presence of an activation or repression domain (Biggin and McGinnis 1997; Yamamoto et al. 1998; Fry

and Farnham 1999; see section 3.5.3). Some transcription factors require other bound proteins to function; others interact synergistically, producing a much stronger effect on transcription in combination than alone (Sauer, Hansen, and Tjian 1995; Thanos and Maniatis 1995). More dramatically, some transcription factors function as either activators or repressors in different contexts. This can happen in at least four ways: (1) *activation and repression domains may be present in the same protein* (e.g., Dorsal: Flores-Saaib, Jia, and Courey 2001); (2) *a protein may interact with different partners which contain distinct interaction domains* (e.g., runt: Wheeler et al. 2000; many homeodomain proteins: Knoepfler and Kamps 1995 and Pinsonneault et al. 1997; many nuclear receptor proteins: Benecke, Gaudon, and Gronemeyer 2001); (3) *any DNA-binding protein can act as a repressor if it masks the binding site of a transcriptional activator,* an effect that does not require a specialized repressor domain; and (4) because transcription factors can influence the expression of other transcription factors, *a transcriptional activator can repress other genes through the intermediate step of activating a repressor*, or vice versa (e.g., RPD: Bernstein et al. 2000).

3.4.8 Transcription Factors Are Not Intrinsically Limited to Specific Developmental or Regulatory Roles

The aspects of organismal phenotype affected by a particular transcription factor are determined by the downstream, or target, genes it regulates (and the genes affected by their activity, and so forth). Unlike metabolic enzymes and structural proteins, whose biochemical activities determine phenotype, there is nothing about a transcription factor that intrinsically links it to a particular aspect of phenotype. Some transcription factors have names that imply dedicated functions (e.g., Eyeless in eye development, APETALA1 in floral patterning), but these proteins have additional regulatory roles unrelated to their eponymous ones. Many transcription factors have extended evolutionary associations with particular developmental processes, most famously Hox proteins with anteroposterior patterning in animals (Gerhart and Kirschner 1997) and MADS-box proteins with floral patterning in plants (Lawton-Rauh et al. 2000). In principle, however, any transcription factor could bind to the promoter of any gene and regulate its expression, so long as an appropriate binding site is present; conversely, a gene's transcription could, in principle, be regulated by any transcription factor. This lack of an obligate connection between a specific transcription factor and a specific aspect of phenotype is evident on both developmental and evolutionary time scales: during development most transcription factors are expressed in several temporally and spatially distinct phases, where they may regulate the expression of different downstream genes that influence completely unrelated aspects of phenotype (Duboule and Wilkins 1998; Davidson 2001), and during the course of evolution the downstream genes regulated by a transcription factor can change dramatically (Keys et al. 1999; Davidson 2001; Wilkins 2002).

3.5 Promoter Function

The transcriptional output of a promoter is not a simple function of which binding sites are present. The relative position, orientation, and nucleotide sequences of these binding sites, as well as the expression profiles of their cognate transcription factors and cofactors, all interact to produce the total transcription profile of a gene. These interactions are complex, nonlinear, and often strongly context dependent. At least for the near term, we lack the ability to predict transcription profiles from sequence inspection.

3.5.1 Transcription Is ''Off'' by Default

Native chromatin is impervious to the RNA polymerase II complex, and even a decondensed basal promoter cannot efficiently direct transcription in the absence of specific transcription factors (Carey and Smale 2000; Courey 2001). Because transcription is ''off'' by default, all promoters contain binding sites for activators of transcription but only some contain binding sites for negative regulators (Arnone and Davidson 1997; Davidson 2001). Although not ubiquitous, repression is common and can be important for modulating the level of transcription, for restricting expression from inappropriate regions, and for adjusting gene expression in response to extracellular signals (Gray and Levine 1996; Shore and Sharrocks 2001). Activating transcription requires decondensing the chromatin surrounding the basal promoter and around some transcription factor binding sites, followed by DNA binding by specific transcription factors capable of recruiting the RNA polymerase II complex onto the basal promoter. In practice, activating transcription at a single locus requires dozens of specific interactions among macromolecules (Thanos and Maniatis 1995; Reinberg et al. 1998; Wolffe 2001) (fig. 1*B*).

3.5.2 Binding Site Position and Orientation Are Functionally Tied More Closely to Nearby Binding Sites than to the Basal Promoter

Some protein-protein interactions depend on precise spacing and relative orientation of binding sites (e.g., Hanes et al. 1994). In particular, steric hindrance and some cases of cooperative binding require binding sites to be in specific positions relative to each other. These interactions involve binding sites that typically lie no farther apart than the size of the proteins that they bind (in practice, up to a few tens of base pairs apart). Some interactions are precisely phased to lie on the same side of nucleosomes (~40-bp multiples) or completely decondensed DNA (~10-bp multiples) (Lewin 2000; White 2001). In contrast, many protein-protein interactions take place through DNA looping and are relatively insensitive to position and orientation. Binding sites that lie more than a few tens of base pairs away from the basal promoter must interact with it via DNA bending or looping, and they often tolerate changes in position relative to the transcription unit (indeed, this was part of the original operational definition of an ''enhancer'': Serfling, Jasin, and Schaffner 1985; Atchison 1988). Binding sites that interact with

chromatin remodeling complexes may also have only moderate functional constraints on position and orientation. This combination of position sensitivity for local interactions and position insensitivity for interactions with effector complexes may underlie the evolutionary origin and maintenance of promoter modules (see section 3.5.4).

3.5.3 The Regulatory Role of a Binding Site Is Often Context Dependent

Some binding sites have discrete functions within promoters, in the sense that when they are occupied by a protein they consistently have the same effect on transcription. In many instances, however, the consequences of transcription factor binding are strongly context dependent (fig. 4) (Biggin and McGinnis 1997; Fry and Farnham 1999; Lemon and Tjian 2000; Courey 2001). (1) *The presence or absence of cofactors is often important*. Many transcription factors interact with transcriptional or chromatin-remodeling complexes through cofactors (see section 3.4.5). For instance, the transcriptional activator CREB requires the cofactor CBP to recruit the RNA polymerase II complex (Shikama, Lyon, and La Thangue 1997), whereas the repressor protein groucho requires one of several cofactors to initiate chromatin condensation (Chen and Courey 2000). (2) *Some binding sites bind different transcription factors under different circumstances*. The consensus binding sequences of many transcription factors overlap (table 4). The effects of different proteins that interact with the same binding site can differ depending on the protein-protein interaction domains they contain. For instance, the locus that encodes the transcription factor CREM in mice produces six distinct isoforms, all of which can recognize some of the same binding site sequences; some isoforms contain protein-protein interaction domains that activate transcription, while others lack these domains and block the activator isoforms from binding (Foulkes and Sassone-Corsi 1992). (3) *Some binding sites are positioned sufficiently near each other that only one protein can bind at a time*. Changes in the relative concentrations of transcription factors that interact with adjacent binding sites can have a significant impact on transcription (see section 3.4.5). (4) *Some transcription factors interact synergistically during or after binding to DNA*. Several cases are known where different transcriptional activators individually have little or no effect on transcription, but in combination they produce a strong effect. For instance, activation of human interferon-beta (IFN-β) transcription requires that several proteins be present, none of which can activate transcription alone (Thanos and Maniatis 1995).

3.5.4 Binding Sites Are Sometimes Organized into Functional Modules

Clusters of nearby transcription factor binding sites sometimes operate as functionally coherent modules (Dynan 1989; Kirchhamer, Yuh, and Davidson 1996; Arnone and Davidson 1997). A module is operationally defined as a cluster of binding sites that produces a discrete aspect of the total transcription profile. A single module typically contains about 6 to 15 binding sites and binds 4 to 8 different transcription factors (Arnone and Davidson 1997; Davidson 2001). Although many promoters contain two or more clearly distinct modules (fig. 2*D* and *E*, 2*H*, 2*J* and *K*, 2*P*), others apparently lack modular organization (fig. 2*A–C* and 2*I*). Modules are often, but not necessarily, physically separated around a locus (compare fig. 2*H*, 2*J* and 2*K* with fig. 2*D*). A single module may carry out one or a combination of the following: (1) *initiate transcription*, often in a highly specific manner such as within a single cell type or region of an embryo; (2) *boost transcription rate* without being able to initiate it; (3) *mediate signals* from outside the cell, by binding a transcription factor that either contains a receptor for a hormone or that is post-translationally modified by a signal transduction system; (4) *repress transcription* under specific conditions or in specific regions or cell types; (5) *restrict the effect of another module* to a single basal promoter through an ''insulator'' function (see section 3.3.6); (6) *selectively ''tether'' other modules*, by bringing them into proximity with a single basal promoter (see section 3.3.6); or (7) *integrate* the status of other modules by influencing transcription differently, depending on what proteins are bound elsewhere (Yuh, Bolouri, and Davidson 1998). The most common term for a promoter module in the literature is an ''enhancer.'' Enhancers were originally defined operationally as segments of DNA capable of elevating transcription in a position-independent and orientation-independent manner (Serfling, Jasin, and Schaffner 1985; Atchison 1988). The term has since been applied much more broadly, to any region of DNA that produces a specific aspect of a transcription profile, sometimes even including regions that repress transcription. Further ambiguities stem from the fact that it is not always possible to assign a single function to a region of a promoter (see section 3.5.3). The terms *enhancer*, *booster*, *activator*, *insulator*, *repressor*, *locus control region*, *upstream activating sequence*, and *upstream repressing sequence*, all refer to various kinds of modules. Although these terms are descriptive of function, they may be misleading if later studies demonstrate multiple functions or context dependence of function. For these reasons, we use the more general term *module* (Dynan 1989; Arnone and Davidson 1997).

3.5.5 Activator and Repressor Modules Are Often Additive in Effect

Experiments often reveal that deleting a single module eliminates a specific aspect of the expression profile without disrupting the remainder (e.g., DiLeone, Russell, and Kingsley 1998; Yuh, Bolouri, and Davidson 1998; Kammandel et al. 1999; Sackerson, Fujioka, and Goto 1999). Conversely, predictable artificial expression profiles can be built by experimentally combining modules from different promoters (e.g., Kirchhamer, Bogarad, and Davidson 1996). These experimental results are the primary basis for the claim that the modularity of promoters contributes to their ''evolvability'' (Stern 2000; Wilkins 2002). In contrast, experimentally deleting insulator, tethering, or integrator modules is epistatic rather than additive (Ohtsuki, Levine, and Cai 1998; Yuh, Bolouri,

and Davidson 1998; Calhoun, Stathopoulos, and Levine 2002).

3.5.6 *Collaboration Among Modules Produces the Total Transcription Profile*

Two aspects of promoter function are reminiscent of analog logic circuits (Yuh, Stathopoulos, and Levine 1998; Davidson 2001; Yuh, Bolouri, and Davidson 2001). (1) Individual modules can function as Boolean (off/on) or scalar (quantitative) elements whose interactions have predictable, additive effects on transcription. Multiple modules are sometimes required to produce a single phase of expression. For instance, the seven stripes of *even-skipped* transcription in the early embryo of Drosophila are controlled by six modules (Sackerson, Fujioka, and Goto 1999; fig. 2*H*). Conversely, a single module may be involved in several different phases of expression. For example, module A in the Endo16 promoter of the sea urchin *Strongylocentrotus purpuratus* (fig. 2*D*) activates transcription in the embryo, synergistically elevates transcription in the larva, and is required for the function of repressor modules (Yuh, Stathopoulos, and Levine 1998; Yuh, Bolouri, and Davidson 2001). (2) Promoters integrate multiple, diverse inputs and produce a single, scalar output: the rate of transcriptional initiation. A familiar analogy is a neuron, which receives input from many sources but whose output is simply how often it fires. In many promoters, signal integration happens at the basal promoter, through specific interactions between bound transcription factors and components of the RNA polymerase II enzyme complex (Latchman 1998; Lee and Young 2000). In some promoters, however, a distinct module may integrate signals from other modules. For instance, module A of the *Endo16* promoter relays the status of the other five modules to the basal promoter (fig. 2*D*; Yuh, Stathopoulos, and Levine 1998; Yuh, Bolouri, and Davidson 2001).

3.6 Gene Networks

All genes are components of immensely complex networks of interacting loci. The binding of a transcription factor to a promoter is one of several physical determinants of gene network architecture. Understanding the organization of gene networks (Lee et al. 2002; Milo et al. 2002) will be necessary for understanding how they evolve (von Dassow et al. 2000; Davidson 2001; Wagner 2001).

3.6.1 *Most Transcription Factors Have Numerous Downstream Target Genes*

A simple calculation demonstrates that most eukaryotic transcription factors must bind to the promoters of many downstream genes. Eukaryotic genomes contain on the order of $0.5–5 \times 10^4$ genes, only a small fraction of which encode transcription factors (table 3 provides a partial list for several species). Because the expression of all genes requires that transcription factors bind to their promoters, and because most promoters contain binding sites for at least five different transcription factors (and often many more), transcription factors must on average interact with the promoters of tens to hundreds of genes. These rough calculations agree with studies that have used experimental approaches to estimate the number of direct downstream targets of a specific transcription factor. In *Drosophila melanogaster*, Ubx isoform Ia alone regulates an estimated 85–170 direct downstream targets (Mastick et al. 1995), whereas *eve* and *ftz* together appear to regulate the majority of genes in the Drosophila genome (Liang and Biggin 1998). The number and identify of direct downstream targets has been assayed by in vivo binding for many transcription factors in *Saccharomyces cerevisiae* (Iyer et al. 2001; Lieb et al. 2001, 2002), and for some transcription factors these analyses have been carried out on cells grown under more than one environmental condition (Ren et al. 2000). Even using conservative criteria for recognizing interactions, these analyses indicate that most transcription factors directly regulate a few percent of the genes in the *Saccharomyces* genome. Genetic networks are therefore highly connected, with each node that is represented by a transcription factor linked to many other nodes. This high degree of connectivity may be responsible in large part for the classical genetic phenomena of epistasis, polygeny, and pleiotropy (Gibson 1996).

3.6.2 *Transcription Is Often Modulated by Feedback Loops*

The expression profile of a gene is a system property, in that it is sensitive to changes in the expression and activity of gene products encoded by many other loci. Thus, even if a mutation in a promoter region alters transcription, the network of functionally interacting genes and gene products may modulate this effect (von Dassow et al. 2000). For instance, a mutation that doubles transcription rate may not result in twice as much protein being produced if there is feedback from the cytoplasm to the nucleus that is sensitive to protein level or to the functional consequences of protein activity (such as accumulation of a particular metabolite). Feedback loops are probably rather common components of gene networks (Lee et al. 2002; Milo et al. 2002) and may mask some functionally significant mutations in promoters.

3.6.3 *A Significant Fraction of the Genome Is Involved in Transcriptional Regulation*

Several large-scale interspecific sequence comparisons have estimated that the number of conserved intergenic nucleotides is similar to the number of conserved coding nucleotides (Shabalina and Kondrashov 1999; Onyango et al. 2000; Bergman and Kreitman 2001; Frazer et al. 2001; Shabalina et al. 2001). This striking result suggests that the number of functional noncoding nucleotides is approximately equal to the number of protein-coding nucleotides, and that approximately half of all functionally or phenotypically penetrant molecular evolution involves noncoding sequences. The hundreds of transcription units encoding general and specific transcriptional factors, chromatin remodeling complexes, and transcription cofactors add to the sequences involved in transcriptional regulation. A substantial fraction of a eu-

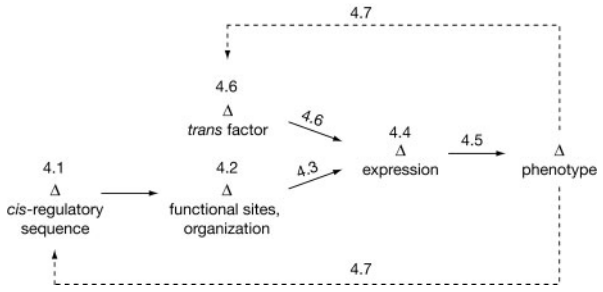

FIG. 5.—Varieties of evolutionary change in transcriptional regulation. The diversity of evolutionary patterns and mechanisms in transcriptional regulation can be organized by their genomic location (*cis* or *trans*) and functional consequence (silent, biochemical, expression, organismal, fitness). The numbers in this diagram refer to sections of the text that discuss that category of evolutionary change.

karyotic genome is devoted to extracting information from itself.

4 The Rich Phenomenology of Promoter Evolution

The literature relevant to promoter evolution is diffuse, and little of it was written by or for evolutionary biologists. Nonetheless, the available information provides a foundation on which to build some initial generalizations. Parallels with, and differences from, the evolution of other regions of eukaryotic genomes are evident. In this section, we document many ways in which promoters and mechanisms of transcriptional regulation have interesting, informative, and varied evolutionary histories. Figure 5 provides an overview of how this section is organized.

4.1 Promoter Sequences Evolve at Different Rates

For reasons that remain poorly understood, coding sequences evolve at markedly uneven rates among lineages and within genomes (Gillespie 1991; Li 1997). Rate variation appears to be a prominent feature of promoter sequence evolution as well. (1) *Long-term conservation*. Similar clusters of transcription factor binding sites are sometimes present in the promoters of orthologous genes of species that diverged up to 10^7–10^8 years ago (e.g., Aparicio et al. 1995; Frasch, Chen, and Lufkin 1995; Gerard, Zakany, and Duboule 1997; Beckers and Duboule 1998; Margarit et al. 1998; Papenbrock et al. 1998; Shashikant et al. 1998; Thompson et al. 1998; Plaza, Saule, and Dozier 1999; Xu et al. 1999; Tümpel et al. 2002). Although sequence similarities in promoters are routinely interpreted as conserved features, another possibility is independent origins of binding sites for the same transcription factor (Cavener 1992; Stone and Wray 2001). Both long-term conservation and parallel origins of binding sites for the same transcription factor suggest constraints on promoter function and imply that stabilizing selection is operating on the gene expression profile. (2) *Rapid divergence*. Promoter sequences can also diverge extensively among even relatively closely related species, and they may include gains and losses of multiple binding sites and changes in the position of regulatory sequences relative to the transcription start site (Wu and Brennan 1993; Takahashi et al. 1999; Wolff et al. 1999; Liu, Wu, and He 2000; Romano and Wray 2003). A comparison of 20 well-characterized regulatory regions in mammals revealed that approximately one-third of binding sites in humans are probably not functional in rodents (Dermitzakis and Clark 2002). Promoter sequence differences may or may not alter transcription (see section 4.3) or organismal phenotype (see section 4.5), depending on genetic background and environmental conditions.

4.2 Functional Changes in Promoters Arise from a Variety of Mutations

Mutations affecting transcription (fig. 6) fall into several distinct classes. (1) *Small-scale, local mutations can modify, eliminate, and generate binding sites and alter their spacing*. Promoter function can be directly altered by the most abundant kinds of mutations: single base substitutions, small indels, and changes in repeat number (e.g., Gonzalez et al. 1995; Shashikant et al. 1998; Segal et al. 1999; Takahashi et al. 2001; Rockman and Wray 2002; Streelman and Kocher 2002). Point mutations can modulate or eliminate transcription factor binding, generate binding sites de novo, or result in binding by a different transcription factor (''transcription factor switching'': Rockman and Wray 2002). Insertions and deletions can change spacing between binding sites, as well as eliminate binding sites or generate new ones (Ludwig and Kreitman 1995; Belting, Shashikant, and Ruddle 1998). Changes in microsatellite structure can affect spacing between binding sites and alter the number of binding sites, sometimes with functional consequences (Trefilov et al. 2000; Rockman and Wray 2002; Streelman and Kocher 2002). (2) *New regulatory sequences can be inserted into promoters through transposition*. This phenomenon has been reviewed extensively (Britten 1997; Kidwell and Lisch 1997; Brosius 1999). For instance: B2 SINEs in *Mus musculus* contain sequences capable of acting as basal promoters (Ferrigno et al. 2001) and some Alu elements in humans contain binding sites for nuclear hormone receptors and exert an influence on transcription (Babich et al. 1999). (3) *Retroposition may assemble new promoters*. Retroposition can create novel genes that are subsequently expressed (e.g., *jingwei* and *sphinx*: Long, Wang, and Zhang 1999; Wang et al. 2002). This process occurs at appreciable frequencies within the genus Drosophila (Bétran et al. 2002). The molecular mechanisms underlying retroposition preclude transfer of the basal promoter and virtually all *cis*-regulatory sequences (the exception being those within exons). Because no gene can function without transcriptional regulatory sequences, it seems likely that novel genes that arise through retroposition either fortuitously insert near existing *cis*-regulatory sequences and come under their regulation without disrupting existing regulatory fuctions or persist long enough that novel *cis*-regulatory sequences arise through transposition, recombination, or small-scale local mutations. Remarkably, novel genes that arise through retroposition are often expressed in tissue-specific patterns

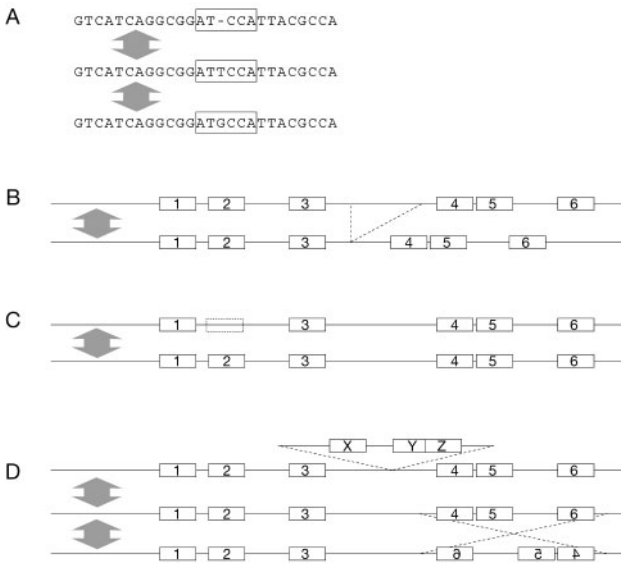

FIG. 6.—Mutations affecting promoter structure and function. (*A*) Modifications in the sequence of a transcription factor binding site. The simplest such changes involve single nucleotide substitutions, insertions, or deletions. Repeat length changes can also affect binding sites. (*B*) Modifications in the spacing of binding sites. These changes can arise in several ways: from insertions and deletions of other genomic segments or mobile elements (as shown), from the accumulation of small indels, and from the expansion or contraction of repeats. (*C*) Modifications in the presence or absence of functional binding sites. Binding sites can arise or be lost through local point mutations, insertions, or deletions. (*D*) Large-scale changes in promoter organization. Several kinds of reorganization are possible, just two of which are shown. Mobile element insertions can ''import'' functional binding sites into a promoter (see section 4.2). Small chromosomal rearrangements can alter the orientation or location of clusters of binding sites. The functional consequence of each kind of change can range from no effect on gene expression to altered expression to loss of expression (see section 4.3).

similar to those of a parent locus (Bétran et al. 2002). (4) *Gene duplications may fragment or recombine promoter sequences*. Although analyses of gene duplication typically focus on coding sequences, the associated promoters are clearly also important for gene function. If the breakpoints do not include *cis*-regulatory sequences, then the duplicated copy is likely to be transcriptionally inert in its new location and become a pseudogene even before it accumulates stop codons or frameshifts. If only part of the promoter is duplicated, the transcription profile of the new copy may differ from the original (e.g., *nNOS*: Korneev and O'Shea 2002). In principle, a duplication could also fortuitously combine sequences from two different promoters to create a hybrid *cis*-regulatory region with a novel transcription profile. Gene duplications that persist are frequently followed by divergence in expression (Li and Noll 1994; Gu et al. 2002; Stauber, Prell, and Schmidt-Ott 2002) and may be followed by loss of complementary promoter modules (Ferris and Whitt 1979; Force et al. 1999). (5) *Gene conversion can spread regulatory elements within a gene family*. Examples from humans include growth hormone (Giordano et al. 1997), beta and gamma globins (Chiu et al. 1997; Patrinos et al. 1998), and major histocompatibility complex (MHC) genes (Cereb and Yang 1994). Gene conversion is an ongoing process in RNA polymerase I–transcribed genes (which encode the 40S pre-rRNA that is processed to form 18, 5.8, and 28S rRNA) including associated transcriptional regulatory sequences, but not among the more heterogeneous RNA polymerase III-transcribed genes (White 2001). (6) *Sequences that have no prior function in regulating gene expression can become fortuitous promoters*. In one case, gene duplications resulted in a former exon functioning in transcriptional regulation (*Sdic*: Nurminsky et al. 1998). A second example involves functional transcription factor binding sites within an exon (*nonA*: Sandrelli et al. 2001). These ''hopeful monster'' promoters demonstrate that rare events can assemble functional *cis*-regulatory sequences from seemingly unpromising material.

4.3 Changes in Promoter Sequence Differ Widely in Their Effects on Transcription

Relatively little information exists about the functional consequences of naturally occurring differences in promoter sequences. A few studies have directly examined the biochemical impact of sequence differences on protein binding (e.g., Ruez, Payre, and Vincent 1998; Singh, Barbour, and Berge 1998; Wolff et al. 1999; Shaw et al. 2002). Most of what we know about functional consequences, however, comes from cases in which the resulting transcription profile has been examined (e.g., Ross, Fong, and Cavener 1994; Odgers, Healy, and Oakeshott 1995; Tournamille et al. 1995; Belting, Shashikant, and Ruddle 1998; Indovina et al. 1998; Ludwig, Patel, and Kreitman 1998; Wang et al. 1999; Romano and Wray 2003). In some of these cases, specific promoter sequence differences are correlated with phenotypic consequences; in most, however, the presence of multiple sequence differences makes it difficult to infer the precise basis for evolutionary changes in transcription. Divergence in promoter sequence and transcription profile are often poorly correlated: very similar promoters can produce substantially different transcription profiles (e.g., parvoviruses: Storgaard et al. 1993; *TNFA* in primates: Haudek et al. 1998; *MMSP* in diptera: Christophides et al. 2000), whereas highly divergent promoter sequences can produce very similar transcription profiles (e.g., *runt* in Drosophila: Wolff et al. 1999; *brachyury* in ascidians: Takahashi et al. 1999; *yp* in Drosophila: Piano et al. 1999; *Endo16* in sea urchins: Romano and Wray 2003).

The latter situation is not unusual (for additional examples, see section 4.7). Indeed, many changes in promoter sequence do not alter transcription within the limits of experimental assays. Sequence changes might be functionally silent for several reasons. (1) *Substitutions and indels between transcription factor binding sites may not affect DNA-protein interactions*. It is probably generally true that nucleotides within binding sites are more functionally constrained than those that lie between binding sites. However, it is difficult to rule out the possibility that a supposed nonbinding site nucleotide might in fact be part of an unrecognized binding site (see section 5.2). Furthermore, small indels that do not directly involve binding sites may disrupt protein-protein interactions by placing proteins on opposite sides of the DNA

helix or by changing the spacing of binding sites (see section 3.5.2). (2) *Changes in spacing between distant binding sites will be neutral in many cases*. Interactions among proteins associated with binding sites more than ~50 bp apart are probably mediated by DNA bending or looping, which may to a large degree be insensitive to differences in spacing. (3) *Some within-consensus nucleotide substitutions in binding sites may be functionally neutral*. Certain changes in binding site sequence can preserve a particular DNA-protein interaction (see section 3.3.3). Not all such changes will be neutral, however, as binding kinetics may differ, in turn altering transcription (see section 3.3.4). Very little is known about the evolution of binding site consensuses, so sequence comparisons alone may be a poor guide as to which nucleotide substitutions within binding sites are likely to be functionally neutral. (4) *Eliminating an entire binding site may be functionally neutral*. Many promoters contain multiple copies of the same binding site, raising the possibility of functional redundancy. Cases of probable binding site turnover (Ludwig and Kreitman 1995; Hancock et al. 1999; Piano et al. 1999; Liu, Wu, and He 2000; Dermitzakis and Clark 2002; Scemama et al. 2002) may have been possible because of functional redundancy. Nevertheless, multiply represented binding sites within the same promoter are not always functionally redundant (see section 5.5).

4.4 Gene Expression Profiles Evolve Frequently and in Diverse Ways

The literature of comparative gene expression has emphasized similarities and generally interpreted them as conserved features (DeRobertis and Sasai 1996; Holland and Holland 1999; Carroll, Grenier, and Weatherbee 2001). When comparing distantly related taxa, however, similarities in gene expression are often outweighed by apparently nonhomologous features (Wray and Lowe 2000; Davidson 2001; Wilkins 2002). The abundance of population-level variation in promoter function (see section 2.3) means that expression differences could evolve quite rapidly under some conditions, and indeed substantial differences in gene expression can exist even between recently duplicated genes (Gu et al. 2002) or closely related species (Parks et al. 1988; Ross, Fong, and Cavener 1994; Swalla and Jeffery 1996; Grbic, Nagy, and Strand 1998; Kissinger and Raff 1998; Brunetti et al. 2001; Ferkowicz and Raff 2001). Several functional classes of evolutionary change in gene expression are evident. (1) *Changes in timing of gene expression*. Temporal changes have been documented from many taxa (e.g., Dickinson 1988; Wray and McClay 1989; Swalla and Jefferey 1996; Kim, Kerr, and Min 2000; Skaer, Pistillo, and Simpson 2002). Heterochronies are a common pattern of anatomical evolution (McKinney and McNamara 1991), and must, at some level, involve heritable changes in the timing of gene expression. (2) *Changes in spatial extent of gene expression*. Many studies have found interspecific differences in the spatial extent of regulatory gene expression (e.g., Schiff et al. 1992; Abzhanov and Kaufman 2000; Brunetti et al. 2001; Scemama et al. 2002). Such changes are of particular interest when they affect regulatory genes, because of the relatively direct consequences for body proportions, organ size and number, and a great many other anatomical features (see section 2.2). (3) *Changes in level of gene expression*. Evolutionary differences in transcription rate have also been documented (Regier and Vlahos 1988; Crawford, Segal, and Barnett 1999; Wang et al. 1999). Such comparisons have been simplified with the advent of microarray technologies (e.g., Jin et al. 2001; Schadt et al. 2003). Because of this approach, we know more about differences in transcript abundance than any other kind of evolutionary change in gene expression. (4) *Changes in responsiveness of gene expression to external cues*. Evolutionary changes in transcriptional responses to physiological status, environmental conditions, and pheromones have also been documented (Brakefield et al. 1996; Cooper 1999; Abouheif and Wray 2002). Such changes are a necessary component in the evolution of polyphenism and phenotypic plasticity and are therefore of considerable ecological interest. (5) *Sex-specific expression*. Evolutionary changes in differential gene expression among sexes have been documented (Schiff et al. 1992; Saccone et al. 1998; Christophides et al. 2000; Kopp, Duncan, and Carroll 2000). Microarray surveys suggest that populations can harbor variation in which genes are expressed in a sex-specific manner (Jin et al. 2001). (6) *Gains and losses of particular phases of gene expression*. In multicellular organisms, many genes are expressed in a succession of spatially and temporally distinct phases during the life cycle (for examples see Gerhart and Kirschner [1997]; Carroll, Grenier, and Weatherbee [2001]; and Davidson [2001]). A gene whose expression requires a particular transcription factor during a specific phase of expression may be ''abandoned'' by that regulator if it is no longer expressed in the appropriate region. Examples include several independent losses of patterning roles for homeodomain transcription factors in arthropods (Dawes et al. 1994; Falciani et al. 1996; Grbic, Nagy, and Strand 1998; Mouchel-Vielh et al. 2002). Conversely, a new regulatory linkage may be established if a promoter acquires a binding site for a different transcription factor, a process known as *recruitment* or *co-option* (Duboule and Wilkins 1998; Wilkins 2002). Many likely cases have been identified (Lowe and Wray 1997; Saccone et al. 1998; Keys et al. 1999; Brunetti et al. 2001; reviewed in Wilkins 2002). Evolutionary gains and losses of particular phases of gene expression may be facilitated by the modular organization of promoters (see section 2.1).

4.5 Changes in Gene Expression Differ Widely in Their Effects on Organismal Phenotype

Promoter function has both a biochemical phenotype, the gene expression profile, as well as an organismal phenotype, involving features such as anatomy, physiology, life history, and behavior. These biochemical and organismal effects are evolutionarily dissociable to some extent, because some changes in gene expression appear to have no consequence for organismal phenotype. Such changes in gene expression are analogous to conservative amino acid replacements in a protein (table 5), many of

Table 5
Functional Categories of Nucleotide Substitution

Probable Impact	Coding Sequence[a]	Promoter Sequence
Neutral	Synonymous nt substitution	Nonbinding site substitution
Low	Conservative AA replacement	Within consensus substitution
Medium to high	Nonconservative AA replacement	Nonconsensus substitution
Loss-of-function	Frameshift or stop codon	Deletion of basal promoter or key activator binding site

[a] AA = amino acid.

which are likewise thought to have no impact on organismal phenotype (Kimura 1983; Gillespie 1991). Several cases are known where the timing or spatial extent of gene expression differs among species without any obvious phenotypic consequence (e.g., *gld* in Drosophila: Schiff et al. 1992; Ross, Fong, and Cavener 1994; *Cy* gene family in sea urchins: Fang and Brandhorst 1996; Kissinger and Raff 1998; *msp130* in sea urchins: Wray and Bely 1994).

Although it may be difficult to demonstrate beyond any doubt that a particular difference in transcription is phenotypically silent, the opposite case is easier to establish. Differences in gene expression have been linked to diverse aspects of organismal phenotype, including: (1) *anatomy* (Burke et al. 1995; Averof and Patel 1997; Stern 1998; Wang et al. 1999; Lettice et al. 2002); (2) *physiology* (Abraham and Doane 1978; Matsuo and Yamazaki 1984; Dudareva et al. 1996; Sinha and Kellogg 1996; Stockhaus et al. 1997; Segal, Barnett, and Crawford 1999; Lerman et al. 2003); (3) *behavior* (Trefilov et al. 2000; Caspi et al. 2002; Enard et al. 2002*b*; Fang, Takahashi, and Wu 2002; Hariri et al. 2002; Saito et al. 2002); (4) *disease susceptibility* (Tournamille et al. 1995; Shin et al. 2000; Bamshad et al. 2002; Meyer et al. 2002); (5) *polyphenism* (Brakefield et al. 1996; Abouheif and Wray 2002); and (6) *life history* (Allendorf, Knudsen, and Phelps 1982; Allendorf, Knudsen, and Leary 1983; Anisimov et al. 2001; Streelman and Kocher 2002).

4.6 Mutations in *trans* Can Alter Transcription in Several Ways

The genetic basis for an observed difference in the expression of a particular gene in some cases does not reside in *cis*, but rather within one of the loci encoding transcription factors that interact with it. Three classes of mutations can underlie these *trans* effects. (1) *Mutations affecting the expression profile of an upstream transcription factor*. Numerous experiments demonstrate that this *trans* effect is pervasive: manipulating the expression of a transcription factor typically alters the expression of its downstream targets (Gilbert 2000; Alberts et al. 2002). Although many evolutionary differences in the expression profiles of transcription factors are known (see section 4.4), few studies have investigated the effect of these changes on the transcription of downstream targets. Indirect evidence of an evolutionary role comes from phenotypic correlates of interspecific differences in transcription factor expression (e.g., Burke et al. 1995; Averof and Patel 1997; Stern 1998; Beldade, Brakefield, and Long 2002) and from expression assays that test regulatory sequences of one species in another (e.g., Manzanares et al. 2000). (2) *Mutations affecting the DNA-binding domain of an upstream transcription factor*. Amino acid substitutions in DNA binding domains of transcription factors can affect the expression of downstream genes (e.g., Conlon et al. 2001; D'Elia et al. 2002) and produce phenotypic consequences (Brickman et al. 2001). Such changes are apparently relatively rare, as the amino acid sequences of DNA binding domains are usually highly conserved (Duboule 1994; Latchman 1998). Nonetheless, variants are sometimes found within populations (e.g., Brickman et al. 2001). Interspecies gene-swapping experiments support this view: in a surprising number of cases, a vertebrate gene encoding a transcription factor can restore a somewhat wild-type phenotype to a fly that is homozygous for a null allele of the orthologous gene (e.g., Lutz et al. 1996; Gerard, Zakany, and Duboule 1997). Few gene swaps rescue phenotype perfectly and some fail almost completely, however, which may be due in part to changes in DNA binding specificity. (3) *Mutations affecting the presence or sequence of a protein-protein interaction domain in an upstream transcription factor*. Again, experiments provide evidence of this third class of *trans* effects on transcription (Hope and Struhl 1986; Dawson, Morris, and Latchman 1996). Functional changes in protein-protein interaction domains have evolved in Hox transcription factors within the Arthropoda (Galant and Carroll 2002; Ronshaugen, McGinnis, and McGinnis 2002) and in serum response transcription factors between arthropods and chordates (Avila et al. 2002), while an evolutionary difference in a phosphorylation site has evolved in the FOXP2 transcription factor along the lineage separating humans from the other great apes (Enard et al. 2002*b*). Sequence comparisons suggest that amino acid substitutions in protein-protein interaction domains can evolve rapidly under positive selection (Sutton and Wilkinson 1997; Barrier, Robichaux, and Purugganan 2001). All three classes of *trans* effects mentioned above are likely to be highly pleiotropic because of the large number of downstream target genes that would be affected (see section 3.6.1).

4.7 Many Modes of Selection Operate on Promoter Sequences

The classic modes of natural selection that operate on coding sequences and morphology also operate on promoter sequences (also see section 2.4). (1) *Negative (purifying) selection*. Many deleterious promoter alleles have been identified in humans, involving a wide range of genes and phenotypic consequences (summarized in Cooper 1999). Cases of long-term conservation of binding sites (see section 4.1) suggest persistent negative selection. (2) *Positive selection*. Some promoter alleles appear to be under directional selection (*FY*: Hamblin and DiRienzo 2000; *P450*: Daborn et al. 2002; *hsp70*: Lerman et al. 2003). (3) *Overdominant selection*. Likely cases include

some histocompatibility loci in humans and mice (Guardiola et al. 1996; Cowell et al. 1998). Reasons for overdominant selection on coding sequences of these loci are reasonably well understood, and transcription profiles should be under selection for variation in the cell type in which they are expressed (Guardiola et al. 1996). Other possible cases include some β-thalassemias in humans (Kazazian 1990), anthocyanin pigment synthesis in maize (*r* locus: Li et al. 2001), and dispersal behavior in rhesus macaques (*serotonin transporter*: Trefilov et al. 2000). (4) *Balancing selection.* Environmental heterogeneity within the range of a single species can result in local adaptation and balancing selection. Cases are known from a teleost (*LDH*: Crawford, Segal, and Barnett 1999; Segal, Barnett, and Crawford 1999) and from humans (*CCR5*: Bamshad et el. 2002). (5) *Stabilizing selection.* When binding sites within a promoter differ but the resulting expression profile is unchanged, stabilizing selection may be operating. This situation appears to be relatively common, with examples now known from diverse metazoans and genes (*ADH*: Wu and Brennan 1993; *Esterase-5B* and *-6*: Odgers, Healy, and Oakeshott 1995; Tamarina, Ludwig, and Richmond 1997; *unc-116*: Maduro and Pilgrim 1996; *even-skipped*: Ludwig, Patel, and Kreitman 1998; Patel et al. 2000; *yolk protein*: Piano et al. 1999; *runt*: Wolff et al. 1999; *brachyury*: Takahashi et al. 1999; *achaete-scute* complex: Skaer and Simpson 2000; *Hoxb2*: Scemama et al. 2002; mating-type loci in budding yeast: Sjostrand, Kegel, and Astrom 2002; *Endo16*: Romano and Wray 2003). (6) *Compensatory selection*: An interesting case in humans involves a hypomorphic allele within the coding sequence of *CFTR* that causes cystic fibrosis. Some haplotypes contain a second mutation within the promoter that adds a third Sp1 binding site, elevating transcription and resulting in an improved prognosis (Romey et al. 1999, 2000). The third Sp1 site never occurs in haplotypes that produce wild-type protein, suggesting that it may be under positive selection as a result of its compensatory effect.

5 Challenges in Studying Promoter Evolution

The structure and function of promoter sequences are profoundly different from those of coding sequences (table 6). These differences impose nontrivial challenges for studying the evolution of transcriptional regulation.

5.1 Coding and Regulatory Sequences Differ in Structure and Function

Coding sequences have a regular, direct, precise, and easily interpreted relationship with their proximate (biochemical) phenotype, a specific sequence of amino acids. In contrast, promoters have an idiosyncratic, indirect, nonlinear, and context-dependent relationship with their proximate phenotype, a particular transcription profile. Furthermore, the transcription profile generated by a promoter depends on other loci that encode transcription factors that bind to it, and on the loci encoding the transcription factors that regulate these immediate upstream regulators, and so forth. This regress transcends generations, in that maternally loaded transcription factors or their mRNAs are required to activate early zygotic gene expression. Environmental influences on gene expression add a further layer of complexity. Although the amino acid sequence of a protein rarely changes during the course of development or in response to environmental conditions, the transcription profile of most genes is modulated during the life cycle and in response to changing external conditions. (Even when differential splicing produces distinct protein isoforms from the same locus under different circumstances, the relationship between DNA and protein sequence remains direct.)

It is important to recognize that sequence data alone cannot reveal the organization of binding sites within a promoter; nor can they show what proteins bind to them, or how they function, or what transcription profile they generate. This is partly a matter of missing information: for instance, the full matrix of binding sequences is not yet known for most transcription factors, even in well-studied species. But it is mostly an inescapable consequence of the way transcription is regulated: many potential binding sites have no influence on transcription in vivo, sequences essential for transcription always reside both *cis* and *trans*, and transcription can be strongly influenced by genetic background, physiological status, and environmental conditions.

5.2 Identifying Functional Binding Sites Requires Biochemical Data and in Vivo Functional Assays

Because the sequences bound by transcription factors are short and imprecise (see section 3.3), literally hundreds of potential binding sites lie near every locus. Only a fraction of these binding sites actually influence transcription (Latchman 1998; Weinzierl 1999; Li and Johnston 2001; Lee et al. 2002). Potential binding sites may not function for a variety of reasons (see section 3.5.3; fig. 4). Which sites actually influence transcription, and are therefore possible targets of selection, can only be determined experimentally. Biochemical characterizations can identify binding sites precisely and are the only way to determine whether consensus sequences differ among species. The most common methods are footprinting and mobility shift assays (Carey and Smale 2000). Because these assays are carried out in vitro, they cannot reveal the influence of chromatin modulation on protein binding or transcription. Assays of in vivo binding sites (Walter and Biggin 1996; Ren et al. 2000; Iyer et al. 2001; Lee et al. 2002) provide a more accurate representation but are technically more demanding and undercount real binding sites. The only definitive means of identifying a binding site with a role in regulating transcription is to modify its sequence and assay transcription in vivo, typically by transient or stable transformation with a reporter gene (see section 5.5). All of the methods mentioned above require considerable effort when used to test a potential binding site at multiple phases of the life cycle and under a variety of environmental conditions. In practice, most promoters have only been searched for potential binding sites at a restricted phase of the life cycle and under uniform culture conditions. For this reason, experimentally verified binding sites are

Table 6
Structural and Functional Differences in Coding and Promoter Sequences for a Protein-Coding Locus

	Coding	Promoter
Physical boundaries	*Defined by sequence*	*Not defined by sequence*
Start	ATG (sometimes multiple)	None
End	TAA, TAG, or TGA	None
Internal[a]	[C/A]AG | GU[A/G]AGU $(U/C)_nNAG$ | G/A	None
Physical organization	*Discontinuous, colinear*	*Discontinuous, nonlinear*
Physical unit	Exon	Module[b]
Typical unit size	~ 20–2,000 bp	~ 200–2,000 bp
Number of units	1–10 (rarely more)	1–10 (rarely more?)
Organization and function colinear	Yes	No
Relative order of units	Consistent	Not consistent
Modules correspond to functions	Sometimes	Often
Functional organization	*Direct, local*	*Indirect, distributed*
Direct functional output	Protein sequence	Transcription profile
Unit of information	Codon	Binding site
Number of units	~150–1,000 (rarely more)	~6–60 (more?)
Information content	0.3–2.0 kb	0.08–0.8 kb[c]
Spacing between units	Doesn't matter	Sometimes matters
Mapping	Precise (1 AA)[d]	Imprecise (>1 TxF)[d]
Degeneracy	Precise (same AA)	Imprecise (different TxF)
Consequence	Qualitative (1 codon: 1 AA)	Quantitative (level of Tx)[d]
Order of units	Usually matters	Sometimes matters
Genetic basis	*cis* only	*cis* and *trans* required

[a] Type I introns; other splice junction sequences exist.
[b] Cluster of transcription factor binding sites (also known as enhancer, UAS, etc.).
[c] *cis*-regulatory sequences only; additional information necessary for transcription is encoded in the sequences and expressions profile of *trans* regulators and, in some cases, in the environment.
[d] AA = amino acid; Tx = transcription; TxF = transcription factor.

nearly always an underestimate, and the physical extent of a promoter is rarely well defined.

The resulting difficulties for studying promoter evolution are substantial. (1) *Information about promoter structure is almost always incomplete*. Few promoters have been subjected to thorough searches for binding sites, and some binding sites probably remain unidentified even in carefully studied cases. Information about the functional consequences of binding site differences among species is limited to just a few cases (e.g., Singh and Berger 1998; Wolff et al. 1999; Shaw et al. 2002). (2) *The information that does exist is almost always biased*. Because of the way promoter function is typically studied, some kinds of binding sites are naturally less likely to be discovered. These include binding sites that mediate responses to physiological status and environmental conditions (because most assays are carried out under uniform conditions), binding sites that act at restricted times during the life cycle (because typically only part of the life cycle is assayed), and binding sites of weak effect (because of assay insensitivity). In addition, most studies measure either quantitative or spatial aspects of transcription and some ignore temporal changes; as a result, the binding sites that are identified are often biased with respect to their effects on time, space, and level of transcription.

Because empirical validation of binding sites is laborious, attempts have been made to increase the reliability of informatic approaches to binding site identification. We discuss here a few of the many approaches which have been developed (for additional information, see Hardison [2000]; Stormo [2000]; Ohler and Niemann [2001]; and Markstein and Levine [2002]). Most informatic approaches apply either to a specific locus or to a complete genome. In the former category are programs that use databases of known binding site matrices to scan a sequence for potential binding sites (e.g., TRANSFAC: Wingender et al. 2001; EPD: Praz et al. 2002; PlantCARE: Lescot et al. 2002; SCPD: Zhu and Zhang 1999). However, many of the potential binding sites these programs identify have no biological function and are simply spurious matches to a binding site (see previous paragraphs and section 3.3.4). A complementary approach involves comparisons with the homologous chromosomal region from other species, a method known as ''phylogenetic footprinting.'' The rationale is that nucleotides within binding sites are more likely to be conserved by natural selection. This method can successfully identify previously unknown binding sites (Loots et al. 2000; Wasserman et al. 2000; Yuh et al. 2002). The effectiveness of this method is limited, however, because nucleotides can be conserved by chance, because real binding sites can turn over even when the transcriptional output is maintained, and because some aspects of transcription are species-specific (e.g., Ludwig, Patel, and Kreitman 1998; Dermitzakis and Clark 2002). The first problem leads to false positives, whereas the second and third generate false negatives. When a complete genome sequence is available, several additional methods can be applied to identify unknown binding sites. These algorithms rely on large data sets to identify overrepresented sequence motifs (e.g., Sinha and Tompa 2002),

clusters of binding sites (e.g., Berman et al. 2002; Rebeiz, Reeves, and Posakony 2002), or correlations with expression data (e.g., Birnbaum, Benfey, and Shasha 2001). For all of these methods, both false positives and false negatives remain a significant issue. Although methods for informatic detection of binding sites are becoming more sophisticated, for now the results are best viewed as a starting point for empirical validation rather than as a definitive identification of transcription factor binding sites.

5.3 Identifying the Complement of Transcription Factors that Interact with a Binding Site Requires Biochemical Data

A binding site may be occupied by different transcription factors (or by none) at different times or places during development (see section 3.5.3; see also table 4), with distinct functional consequences (Fry and Farnham 1999; Lemon and Tjian 2000; Courey 2001). The extent of ''binding site sharing'' by different transcription factors remains poorly understood, because nearly all functional studies of promoters in multicellular organisms have examined a single phase of the life cycle (typically embryos or differentiated cells in culture) under uniform culture conditions.

Overlapping binding site specificities have important implications for evolutionary studies: (1) A transcription factor might not influence transcription, even if a consensus binding site for it exists and is known to bind protein. The presence of a binding site is necessary but not sufficient for transcription factor binding. Demonstrating an interaction between a specific regulator and a binding site requires some form of biochemical characterization, the most common of which are ''supershift'' assays and in vivo footprinting (Carey and Smale 2000). (2) A binding site might be occupied by different transcription factors under different circumstances. Indeed, the protein bound most of the time may not be the one whose consensus recognition motif is the closest match. Recognizing cases of varied binding site occupancy requires testing nuclear extracts across developmental stages, among cell types, and under diverse environmental conditions using supershift assays or in vivo footprinting. (3) The protein that occupies a binding site can change evolutionarily. Likely cases of ''transcription factor switching'' have been identified within humans (Rockman and Wray 2002). In addition, an interaction that has been biochemically validated in one species may not occur in another, even if the sequence of the binding site is perfectly conserved. Demonstrating a conserved or altered protein-DNA interaction requires comparative biochemical data.

5.4 Comparisons of Promoter Sequence Are Not Always Straightforward

Once functional binding sites have been mapped, the next step is identifying homologous binding sites among species or alleles. Promoter sequences can usually be aligned rather easily within a species, although binding sites that fall within repeats can be problematic. In the most straightforward interspecific comparisons, potential binding sites that occupy similar positions, spacing, and orientation relative to the start site of transcription and relative to each other are likely to be homologous. Complications can arise for a variety of reasons: binding site spacing is often functionally unconstrained (see section 3.3.5), transposition can introduce similar binding sites (see section 4.2), and random point mutations can generate new binding sites at an appreciable frequency (Stone and Wray 2001). (We use the term *homologous* in its usual, phylogenetic, sense to denote the hypothesis that a binding site is present in two living species because it was present in their latest common ancestor and has persisted since. Sequence similarity, in contrast, is simply an observation, and can be due to either homology or homoplasy.)

Once homologous binding sites have been identified, routine methods of comparative analysis can be applied to polarize character state transformations, identify reversals and parallel transformations, and reconstruct ancestral states. Most published comparisons of promoter structure involve just two species, with the emphasis typically on identifying conserved binding sites. By surveying more taxa and incorporating functional data, it becomes possible to identify origins, losses, and turnover of binding sites. As with all comparative analyses, dense phylogenetic sampling provides a more robust understanding of evolutionary transformations within promoters, particularly in cases of rapid sequence divergence.

5.5 Promoter Function Can Only Be Determined Experimentally

The only way to determine the expression profile produced by a promoter haplotype is to assay it in vivo, in its normal chromosomal and cell biological contexts. This is most easily accomplished by examining the spatial and temporal distribution of transcripts using in situ hybridization, RNA gel blots, or quantitative PCR. Because even small differences in promoter sequence can alter transcription (see section 4.2), interpreting the functional consequences of such differences among alleles or between species also requires assaying transcription.

Similarly, the only way to understand the contribution of specific sequence differences within a promoter is to carry out comparative functional tests. The most common kind of experiment involves coupling a test regulatory region to a reporter gene whose product is easily detected, and then placing this construct in embryos or cells where it is exposed to the array of transcription factors encountered by the endogenous promoter (Carey and Smale 2000). Further experiments, such as testing the consequences of nucleotide substitutions within a specific binding site, deleting a binding site, altering spacing or orientation between binding sites, or testing restricted portions of the promoter can be immensely informative. Although such experiments are laborious, they provide almost the only reliable information about binding site function. Fortunately, expression assays are feasible in a growing number of organisms. For comparative analyses, it is important to carry out reciprocal functional tests, because transcription

is a product of both *cis* and *trans* sequences. This, too, is now possible in some taxa (Tümpel et al. 2002; Romano and Wray 2003).

Comparative experimental tests, although unusual in the literature, are necessary for analyzing the evolution of promoter function. Issues of particular interest include the following. (1) *Identifying changes in a binding site function.* A binding site whose sequence is conserved between two species may nonetheless function differently in them, because the transcription factors and cofactors that interact with it are expressed differently or because an adjacent binding site for a cofactor has changed. (2) *Determining the function of multiply-represented binding sites.* Multiple binding sites for the same transcription factor within a single promoter (fig. 3) may be functionally redundant, additive, or synergistic (e.g., Small, Blair, and Levine 1992; Yuh and Davidson 1998, 2001), with distinct consequences for selection (Ludwig 2002). (3) *Understanding the functional significance of binding site organization.* The position, spacing, and orientation of individual binding sites in some cases matters a great deal and in other cases not at all (see 3.5.2), again with distinct consequences for selection. (4) *Determining when, where, and under what conditions a binding site functions.* Although some binding sites may function continuously and ubiquitously, most probably do so only during part of the life cycle, in certain cell types, or in response to particular environmental conditions. Binding site function may also be context dependent, changing under different circumstances (see section 3.5.3). (5) *Identifying the genetic basis for a known difference in transcription.* Interspecific differences in transcription profiles might be due to changes in *cis* or *trans* (see sections 4.2 and 4.6). The functional consequences of changes in *cis* can be identified by means of in vivo expression assays (examples reviewed in Paigen [1989] and Cavener [1992]).

The difficulty of carrying out comparative expression assays imposes severe practical constraints on analyses of promoter evolution. (1) *In general, it will be more difficult to obtain comparative information on the proximate function of promoter sequences than of coding sequences.* Characterizing promoter function involves techniques that are labor intensive and unfamiliar to most molecular evolutionists. Yet without this information, it is difficult to interpret comparative sequence data meaningfully. (2) *At least for the near term, comparative information on binding site function will remain limited.* Few promoters have been analyzed biochemically or functionally in more than one species, and even in these cases analyses have been limited to a fraction of the complete *cis*-regulatory region. (3) *Predicting the proximate functional consequences of mutations in promoter sequences will be more difficult than in coding sequences.* Because promoters lack a general organizing ''code,'' because the function of a binding site can be strongly context-dependent (see section 3.5.3), and because promoter function depends on the sequences and expression of transcription factors encoded elsewhere, the biochemical phenotypic consequences of sequence differences in a promoter region are very difficult to interpret without functional tests. The relative magnitude of likely functional consequences of a mutation within a promoter can be organized into a very rough rank order, as it can be more precisely within exons (table 5). Overall, however, the considerably less regular structure-function relationship within promoters will make it much more difficult to discern general patterns of sequence evolution.

5.6 Standard Classed Tests of Molecular Evolution Must Be Modified for Analyzing Promoters

Although tests of selection on promoters are not fundamentally different from tests on coding regions, they must be applied with caveats. The major problems arise when applying tests that use classes of nucleotide substitutions to promoter data (e.g., K_a/K_s or McDonald-Kreitman tests; McDonald and Kreitman 1991). These tests classify coding mutations as synonymous or nonsynonymous, and they test for selection under the assumption that synonymous sites evolve neutrally. To apply these tests to promoter sequences, most authors classify promoter mutations as occurring within binding sites or within non-binding-site nucleotides and assume that nonbinding sites are evolving neutrally. However, the functional consequences of mutations in promoters cannot be classified without additional functional data (see section 5.5; Jenkins, Ortori, and Brookfield 1995). More specifically, practical difficulties in identifying binding sites (see section 5.2) mean that most evolutionary analyses of promoters will only be able to rely on functional information from a single species. Binding sites absent in the species for which functional data are available but present in all other species will be missed, while those sites functional in the reference species but not in all other species may mistakenly be considered present in all. Both types of error will result in some sequence differences being classed incorrectly and will degrade the signal-to-noise ratio in tests for selection. It follows that sequence comparisons among the promoters of closely related species, or classed tests that only use data from one species (e.g., Hahn, Rausher, and Cunningham 2002), will generally be the most informative and accurate. Even fewer data from comparative studies are usually available about the functional consequences of sequence differences within a binding site. Only rarely will it be possible to reliably detect the effect of nucleotide differences on changes in binding specificity between species (see section 5.5).

A second problem with tests that rely on classes of sites relates to the mechanism by which binding sites arise and are presumably selected for. Although an excess of nonsynonymous substitutions relative to synonymous substitutions is good evidence for positive selection, it is difficult to imagine a situation in which an excess of binding-site substitutions relative to nonbinding-site substitutions can be interpreted in the same way. This follows from three features of promoters. First, binding sites are sometimes not functionally restricted to a specific position. Individual binding sites may therefore turn over by changing position without positive selection (Ludwig 2002). Second, sequences which have no binding affinity for any transcription factor often need only a single base-pair change in order to become a functional binding site

(Stone and Wray 2001). A single point mutation can therefore establish a new functional site consisting of several nucleotides. Third, most nucleotide substitutions within a binding site modulate or eliminate its function, whereas relatively few mutations will change it into a binding site for a different transcription factor. Rarely will an excess of substitutions within a binding site be a signal of positive selection, because binding sites often simply cease to function after multiple substitutions. None of these three features precludes selection for changes in binding sites, only that they may combine to significantly reduce the ability of classed tests to detect this selection in practice.

All classed tests of section suffer from these problems, but non-classed tests of neutrality (e.g., the Hudson-Kreitman-Aguade test [HKA], Tajima's *D*, Fu and Li's *D*: Hudson, Kreitman, and Aguade 1987; Tajima 1989; Fu and Li 1993) can be interpreted in the standard way. Often a combination of these tests, as well as studies of geographic structure of allele frequencies, may be necessary to detect the action of selection in promoters (e.g., Hamblin and DiRienzo 2000; Bamshad et al. 2002; Fullerton et al. 2002).

6 Hypotheses About the Evolution of Transcriptional Regulation

No general framework exists for understanding, interpreting, and predicting how transcription evolves. In this section, we present an initial attempt at providing such a framework, in the form of testable hypotheses derived from three sources: models of molecular evolution, mechanisms of promoter function, and empirical evidence. Our hope is that these hypotheses will encourage investigators to dig a little deeper into their data to address a broad range of questions about promoter evolution. Thus, we emphasize hypotheses that can be tested with available techniques. The first six categories of predictions are based on the neutral model of sequence evolution (Kimura 1983) and are organized as shown in figure 7; the last four categories address promoter ''design'' principles and macroevolution.

6.1 Promoter Sequences Have Characteristic Evolutionary Dynamics

Because promoters are organized and function differently from other regions of the genome, they are subject to distinct functional constraints. Predictable patterns of sequence evolution should therefore distinguish promoters from sequences that lack a role in transcriptional regulation. (1) *Overall substitution and indel frequencies in promoters should be higher than in coding sequences and lower than in introns and in nonregulatory intergenic regions*. Because most nucleotides in promoters do not affect transcription, most substitutions and many indels should have no functional consequence and should therefore evolve without constraint. These predictions are generally supported empirically (Jordan and McDonald 1998; Jareborg, Birney, and Durbin 1999; Waterston et al. 2002). Exceptions might include the promoters of genes encoding proteins that tolerate many amino acid substitutions, are under balancing selection, or whose introns contain regulatory sequences (e.g., *IL-4* and *IL-13* in mammals: Loots et al. 2000). (2) *Indel size spectrum in promoters should be more continuous than that in coding sequences but similar to that in introns, UTRs, and nonregulatory intergenic regions*. Three factors may contribute to a distinctive evolutionary dynamic of length variation within promoter regions: the lack of a reading frame, the low density of functionally important nucleotides, and the ability of many binding sites to operate in a position-independent manner. Indels should be more common, the frequency spectrum of indel size should not be biased toward multiples of three (as it is in coding sequences), and repeat variation and large indels should be much more common. These patterns are evident in some cases (e.g., *hairy*: Kim 2001). (3) *The order of binding sites and modules within promoters should be less conserved than the order of exons within transcription units*. Codons have an obligate colinearity with the amino acid sequence of their protein product, whereas binding sites and modules within promoters can often function to some extent independently of position and order (see section 3.3.5). Gross organizational changes within promoters should be limited largely by mutation, whereas in coding regions such changes should be limited primarily by functional constraints. Small-scale inversions in promoters can exist within populations (e.g., *IGHA1* in humans: Denizot et al. 2001). (4) *''Module shuffling'' should be relatively common*. ''Domain shuffling'' has been an important part of the evolution of many gene families (Lander et al. 2001). The analogous process of module shuffling within promoters may occur at a higher frequency. Several examples of mobile element insertions that have brought functional binding sites into range of a gene are known (Britten 1997; Kidwell and Lisch 1997). The modular organization of many promoters means that a transcription profile could be dramatically modified in a functionally integrated way.

6.2 Selection Acts Primarily on the Sequence and Spatial Arrangement of Binding Sites

The output of a promoter derives from the nucleotide sequences and spatial arrangement of transcription factor binding sites (see section 3). It follows that sequences that lie between binding sites should be free to vary, at most showing weak biases that reflect mutational processes or weak selection to maintain overall base composition or conformational properties. (1) *Negative selection should operate primarily on nucleotide identity within binding sites*. This is the basic idea underlying ''phylogenetic footprinting'' as a method of identifying binding sites (see section 5.2). Several difficulties beset tests for negative selection on binding sites: some nucleotide positions assumed to be ''non-binding sites'' may in fact be part of binding sites that have not yet been identified, some nucleotide substitutions within known binding sites may be functionally tolerated, and binding sites may turn over within a promoter. Nonetheless, some studies have found evidence for preferential conservation of binding sites

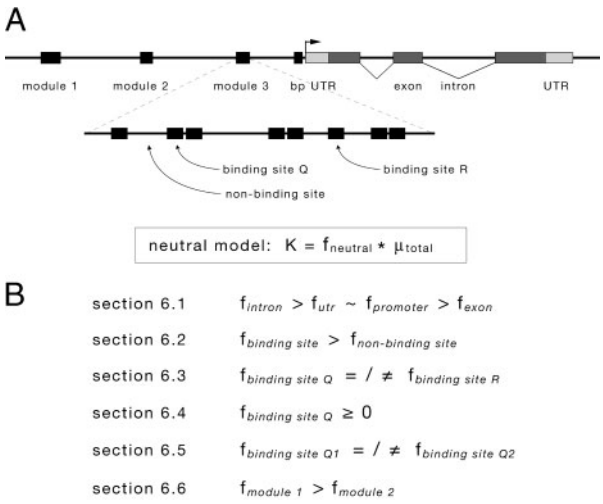

FIG. 7.—A neutral model of promoter evolution. This figure outlines the organization of the first six groups of predictions about promoter evolution (see sections 6.1–6.6), using a neutral model of promoter evolution (i.e., the notion that the rate of evolution at a nucleotide position is inversely related to its functional importance). (*A*) Schematic diagram of a locus, defining the regions referred to below. Using Kimura's (1983) neutral model of sequence evolution, we can model the substitution rate as the fraction of neutral mutations multiplied by the total mutation rate: $K = f\mu_{total}$ ($0 \leq f \leq 1$). (*B*) Relationship of sections 6.1–6.6 to the neutral model. For instance, section 6.1 treats expected differences in patterns of variation among genomic partitions.

(e.g., *teashirt*: Core et al. 1997; *Otx*: Yuh et al. 2002). (2) *Negative selection should operate on the spacing between nearby binding sites*. Protein-protein interactions associated with adjacent binding sites often rely on precise spacing (see section 3.5.2), and small changes in spacing can dramatically affect transcription (e.g., bicoid sites: Hanes et al. 1994; *protein C*: Spek, Bertina, and Reitsma 1999). These functional constraints on length variation are likely to be absent in regions between modules or, more generally, between distantly located binding sites. (3) *Negative selection should eliminate spurious binding sites*. Because they are small, imprecise, and exist for many different transcription factors, binding sites will appear through random mutation at appreciable rates in large populations (Stone and Wray 2001). Where new binding sites interfere with transcription, selection should eliminate them. There is evidence for such selection in prokaryotic genomes, although the strength of selection against such binding sites is estimated to be quite weak (Hahn, Stajich, and Wray 2003).

6.3 Selection Can Discriminate Among Binding Sites Within a Promoter

Some binding sites are more important than others for promoter function. The imprint of different levels of functional constraint among binding sites within a single promoter should be evident in sequence comparisons. (1) *Essential binding sites should evolve relatively slowly*. Functional analyses often reveal that one or a few binding sites are absolutely necessary for activating transcription and that others (usually a greater number) either modulate or have no detectable impact on transcription. Although within-consensus nucleotide substitutions introduce a complication, comparisons should generally reveal lower rates of turnover or loss for essential binding sites. (2) *Binding sites for repressors should evolve faster than those for activators*. There are many ways to repress transcription but relatively few ways to activate it (Latchman 1998; Carey and Smale 2000); furthermore, the consequences of failing to repress transcription may be generally less severe than of failing to activate it (see the next section for more on this point). It follows that the binding sites within a promoter that activate expression may experience stronger negative selection than those that bind repressors. Because binding site function is often context dependent (see section 3.4.7), this pattern may be weak. (3) *Multiply-represented binding sites should evolve faster than unique ones*. In some cases, multiply-represented sites are functionally redundant or each has a minor impact on the overall transcription profile. Thus, selection may tolerate more nucleotide substitutions and turnover of multiply-represented binding sites than unique ones. Some multiply-represented binding sites either function synergistically (e.g., *hunchback*: Ma et al. 1996) or have distinct functions (e.g., *Endo16*: Yuh et al. 2002), which will weaken this prediction. Although several cases of binding site turnover have been identified (Ludwig and Krietman 1995; Liu, Wu, and He 2000; Dermitzakis and Clark 2002; Scemama et al. 2002), direct comparisons of turnover rates in unique versus multiply-represented binding sites have not been made. (4) *Loss of one binding site may be followed by loss of another if their cognate proteins interact*. Binding sites that are occupied by proteins that must interact in order to function will either both be present or both be absent. For promoters with a modular organization, loss of a crucial binding site may lead to eventual loss of the entire module, because the clumped distribution of binding sites into modules is probably a consequence of interactions among the proteins that bind them. In general, the functional interdependence among promoter nucleotides makes these sequences candidates for evolution according to a covarion (Fitch and Markowitz 1970) or fluctuating neutral space (Takahata 1987) model.

6.4 Binding Sites Can Evolve Neutrally

Many point mutations in exons are functionally neutral or near neutral, and some of these are fixed through drift (Kimura 1983; Ohta 1992). The same should be true of promoter sequences. Three categories of sequence change in binding sites, described below, should be effectively neutral (fig. 8). When combined with neutral changes in nucleotides between binding sites (see section 6.2), at least four distinct kinds of neutral sequence evolution should be evident within promoter regions. (1) *Some sequence differences within binding sites should evolve because they do not alter the transcription profile*. Nucleotide substitutions in binding sites that do not alter binding kinetics should not affect transcription, and at the same time, some substitutions that do alter binding kinetics may not affect the transcription profile. Such transcriptionally silent changes within binding sites will accumulate by drift. Because many transcription factors can bind with high specificity to two or more variants

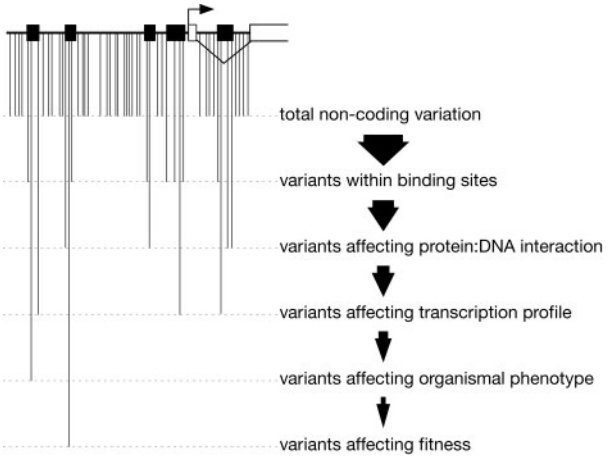

FIG. 8.—Distinct consequences of variants in noncoding sequences. A representation of the variation (or, for interspecies comparisons, fixed differences) in noncoding sequences near a locus. From the total pool of variation, some variants will lie within transcription factor binding sites; a fraction of these will alter protein-DNA interaction; some of these will affect transcription; a subset of these will affect organismal phenotype (anatomy, physiology, behavior, etc.); and some of these will have fitness consequences. These ratios and the kinds of variants that contribute to each are poorly understood for *cis*-regulatory regions.

of a binding sequence (see section 3.4.4), functionally neutral substitutions within binding sites may be relatively common. (2) *Some differences in the complement of binding sites should evolve because they do not alter the transcription profile*. In experimental assays, removing a binding site does not always change the resulting transcription profile, at least within the limits of assay sensitivity (e.g., *SpHE*: Wei et al. 1995). Some evolutionary gains and losses of binding sites may therefore represent transcriptionally neutral changes that were fixed by drift. Functional redundancy provides one avenue: if a new binding site evolves by random walk (Stone and Wray 2001), mutations in an existing site may be tolerated. The new binding site could potentially even bind a different protein, so long as the functional consequence is the same. (3) *Some sequence differences should evolve because they affect transcription but not fitness*. A difference in gene expression need not influence organismal phenotype or fitness. Lack of a fitness consequence for a difference in transcription could lead to the evolution of ''sloppy'' or ''gratuitous'' expression (Gerhart and Kirschner 1997). Neutral variation in gene expression is evolutionarily relevant, because it could later interact with polymorphisms elsewhere in the genome or with changes in the environment to produce phenotypic consequences.

6.5 Selection Can Discriminate Among Variants of a Binding Site

Although many mutations within promoter regions are probably phenotypically neutral, many others certainly are not. Many sequence changes within binding sites affect protein interactions, and this might alter the transcription profile, which could in turn affect fitness. (1) *Specific binding sites may be constrained to a subset of the total binding site matrix of their cognate protein for functional reasons*. The precise affinity of a binding site for a particular transcription factor is sometimes functionally important. In cases where high-affinity and low-affinity variants of the binding site matrix have different phenotypic consequences, negative selection will eliminate variants that bind protein but result in lower fitness. In interspecific comparisons, therefore, some binding sites for a particular protein may show more variation than others. Conversely, specific variants that confer a fitness advantage should be under positive selection (e.g., *ftz*: Jenkins, Ortori, and Brookfield 1995). (2) *Binding sites for transcription factors with many downstream targets should evolve more rapidly than those for factors with few targets*. The consensus binding sites for general transcription factors (TBP, TAFs, etc.; fig. 1*B*) should be fairly broad because they are present in the promoters of many genes: a very narrow consensus would impose a high genetic load because most point mutations in these sites would interfere with binding and thus would likely be deleterious. Thus, binding sites for general factors should exhibit higher levels of variation than binding sites for transcription factors that regulate the specificity of transcription, each of which will be present in fewer genes. This argument can be extended to levels of variation in binding sites for different specific transcription factors: those that that interact with many targets should have broader consensuses than those that interact with only a few targets. (3) *Binding-site specificity of strong activators and repressors should be relatively strict*. Because binding of strong activator or strong repressor proteins is more likely to have a large impact on transcription, selection may operate to narrow the consensus binding sites for these proteins. There are at least three ways, not mutually exclusive, in which this might happen: a requirement for a cofactor that also requires a specific binding site, a relatively large binding site, and a relatively narrow consensus binding matrix.

6.6 Selection Can Discriminate Among Regions and Modules Within a Promoter

Selection should be able to discriminate degrees of functional constraint within a promoter, and the result should be evident as distinct regional patterns of sequence evolution. (1) *Distal regions of large promoters tend to evolve faster than proximal regions*. Binding sites required for initially activating transcription often lie within the first few hundred bases 5′ of the basal promoter, whereas booster, repressor, and tissue-specific modules are often more distant. Physical proximity of activator binding sites to the basal promoter may provide a more reliable or efficient means of initiating transcription (although there are many exceptions). A comparison of human-mouse orthologs found that sequence conservation generally decayed rapidly with distance from the start of transcription (Jareborg, Birney, and Durbin 1999). (2) *Activator modules should evolve more slowly than repressor modules*. The loss of an activator module will, in many cases, be analogous to a stop codon in that it abolishes gene function. In contrast, loss of repressor function is less

likely to be incompatible with gene function. Furthermore, there are many ways to repress transcription, but activation requires a series of specific steps. (3) *Booster modules should evolve somewhat more rapidly than other modules*. Differences in transcript abundance twofold or greater are common within populations (see section 2.3). Promoter modules that modulate transcription level, but that provide no spatial or temporal control, may therefore experience fewer functional constraints on average than most other kinds of modules. (4) *Modules used for multiple phases of expression should evolve more slowly than those used once*. On average, such modules should be under greater functional constraint, which should be apparent as a greater degree of sequence conservation. (5) *Integrator and tethering modules should evolve more slowly than other categories of module*. Promoter modules that function epistatically (see section 3.5.4) may be among the most functionally constrained modules within their respective promoters.

6.7 Structural Complexity in Promoters Reflects Functional Complexity

Genes differ in their functional requirements for regulation: constitutive versus inducible expression, constant level versus modulated level, one versus multiple phases of expression, few inputs versus many inputs, and so forth. These diverse regulatory requirements should be reflected in a similar diversity of functional and organizational complexity in promoters. (By ''complex promoter'' we mean one with relatively many binding sites and regulatory inputs.) (1) *Genes that are constitutively expressed should have simple promoters*. In principle, a promoter that is always and everywhere ''on'' need contain only one binding site for a ubiquitous transcriptional activator. Additional binding sites might be present, however, to add robustness, to set levels of transcription precisely, or to modulate levels in response to extreme conditions such as heat shock. (2) *Regulatory genes expressed early in development should have complex promoters*. The promoters of genes that operate in early embryos drive temporally and spatially precise transcription, despite the fact that pattern formation is ongoing and spatial reference points are not yet well defined. The promoters of these genes often use cooperative protein binding to sharpen boundaries of transcription domains, and this requires additional binding sites. Furthermore, these promoters typically contain binding sites for several positive and negative regulators, as they must integrate multiple spatial and temporal inputs (Arnone and Davidson 1997). (3) *Genes with several distinct expression domains should have complex promoters with modular organization*. Promoters that drive multiphased expression profiles should be more complex, on average, because they respond to many inputs, and that response, in turn, requires interaction with a greater variety of transcription factors. Multiphased expression is very common for genes encoding developmental regulatory proteins. (4) *Genes expressed exclusively in a single differentiated cell type should have simple promoters*. Genes encoding the specialized products of terminally differentiated cells often have relatively simple promoters even though they produce spatially complex expression profiles (e.g., *CyIIa*: Arnone, Martin, and Davidson 1998). The promoters of these genes are typically activated by one or a few tissue-specific transcription factors, and sometimes they lack binding sites for repressors (Davidson 2001). Several so-called master regulators of differentiation are known, including myoD in muscle cells (Yun and Wold 1996) and achaete and neuroD in neurons (Lee 1997). (A corollary of this relationship between dedicated regulators and their downstream targets is discussed in section 6.9.) (5) *Genes that produce more than one isoform should have complex promoters*. Loci that produce multiple isoforms of a protein may generally have more complex promoters, simply because they regulate what is likely to be, on average, a more complex overall expression profile. In addition, alternate transcriptional start sites are often part of the way in which distinct isoforms are generated, adding complexity to such promoters. (6) *Genes with aspects of expression that are contingent on external/extracellular conditions should have more complex promoters*. Signal transduction systems communicate changing conditions in the cytoplasm or at the cell surface to the nucleus, often by phosphorylation or dephosphorylation of a specific transcription factor already present in the nucleus. Contingent regulation of transcription should therefore require additional binding sites for these factors. (7) *Genealogically unrelated genes that are coordinately regulated should share some binding sites*. The promoters of genes that are expressed in similar spatial and temporal patterns share similar functional requirements and should therefore sometimes contain binding sites that evolved independently yet function in a similar manner. Possible cases include insect chorion genes (Mitsialis and Kafatos 1985; Cavener 1992) and vertebrate crystallin genes (Tomarev et al. 1994), among many other examples (Arnone and Davidson 1997; Bernstein, Tong, and Schreiber 2000; Berman et al. 2002). Although few cases of convergence in promoter structure and function have been identified as such, this situation may prove to be common given the ease with which binding sites can be gained (Cavener 1992; Stone and Wray 2001).

6.8 Rates of Promoter Evolution Depend on Many Factors

The diversity of organization and function in eukaryotic promoters (fig. 2; see also the preceding section) should expose them to different modes and degrees of natural selection, which should in turn be reflected in a range of rates and patterns of sequence evolution (see section 4.1). In addition, rates of promoter sequence evolution may be poorly correlated with function, for a variety of reasons. (1) *Promoters containing few binding sites should evolve relatively slowly*. The function of relatively simple promoters may be particularly sensitive to sequence change because they depend on very few proteins for activation and lack multiply-represented binding sites that might confer some functional redundancy. For these and other reasons, the level of functional constraint per binding site may be inversely correlated with the total number of binding sites in a promoter. (2)

Rates of promoter evolution should correlate negatively with codon usage bias at the same locus. A bias in codon usage is generally interpreted to mean that a gene must be translated rapidly at some point in the life cycle (Akashi 2001). Another way to produce protein rapidly is to increase rates of mRNA synthesis. Thus, loci with codon usage bias should have promoters that can direct high rates of transcription. Just as the codons of such genes are biased to a subset of the synonymous possibilities, so too might transcription factor binding sites be restricted to a subset of the binding site matrix that results in high-affinity binding. (3) *A particular mechanism of regulation will sometimes be the target of selection*. In some cases, natural selection may favor a particular mechanism of regulating a conserved transcription profile. For instance, selection might favor stable transcription rates despite environmental perturbations, something that might require additional binding sites beyond the minimum necessary to generate the expression profile under constant environmental conditions. In such cases, the rate of promoter evolution might not be correlated with that of the transcription profile. (4) *Rates of divergence in promoter structure and phenotype should often be uncorrelated*. Within coding sequences, a large discrepancy can exist between the magnitude of a change in genotype and the resulting change in phenotype. For a variety of reasons (see sections 4.3 and 4.6) a similar situation is likely to exist within promoters.

6.9 Some Evolutionary Changes in the Architecture of Gene Networks Are More Likely than Others

The architecture of a gene network (the nature and organization of interactions between genes and gene products) can change during the course of evolution (Wilkins 2002). Some general evolutionary patterns should be evident in how linkages are altered, added (recruitment or co-option), or lost (abandonment). (1) *The genetic basis for an evolutionary change in transcription is likely to reside in* cis. Although a change in transcription could arise in several ways, in practice the genetic basis is likely to reside in the *cis*-regulatory sequence of the downstream gene. This is because most transcription factors regulate the expression of many target genes (see section 3.6.1): thus, a change in the binding specificity or expression profile of a transcription factor will affect the expression profiles of many of its downstream targets, whereas a change in a single binding site for that transcription factor is likely to be much less pleiotropic. Assuming that highly pleiotropic mutations are less likely to become fixed (Fisher 1930), this fundamental asymmetry means that changes in transcription of a given gene are more likely to reside in its promoter than in the amino acid sequences of its upstream regulators (Stern 2000). Protein and microarray expression studies in mouse and humans support this prediction (Klose et al. 2002; Schadt et al. 2003), although a similar study in yeast is equivocal (Brem et al. 2002). (2) *Recruitment of ''top'' and ''intermediate'' regulators should occur more frequently than recruitment of terminal regulators*. Several dramatic evolutionary changes at the top of gene networks are known (sex determination: Hodgkin 1992, Wilkins 2002; embryonic patterning: Stauber, Jackle, and Schmidt-Ott 1999). In each case, the plesiomorphic function of the recruited transcription factor is quite different from the more famous, apomorphic role. Several models have been presented to explain why recruitment of an additional regulator should be tolerated functionally at the beginning or middle of a gene network more often than at their termini (Gehring and Ikeo 1999; Davidson 2001; Wilkins 2002). (3) *Recruitment of a new regulator is more likely to occur for structures that develop in regions and at times where that regulator is already being expressed*. Transcriptional regulators are sometimes expressed in analogous structures, most famously Pax-6 in eyes (Quiring et al. 1994) and Dlx in appendages (Panganiban et al. 1997). Most of these cases appear to involve parallel recruitment rather than mistaken interpretations of comparative anatomy (reviewed by Davidson 2001 and Wilkins 2002). Recruitment to additional regulatory roles is more probable in regions or cell types where the transcription factor is already expressed (Davidson 2001). For instance, *Pax-6* is transcribed in photosensitive neurons throughout the Metazoa, including organisms that lack image-forming eyes. It will therefore be expressed in any organ that contains photoreceptors, and it is more likely to be recruited to additional roles in eye development than transcription factors associated with, for example, muscle cells. (4) *Regulatory linkages to structural genes should be more conservative than those between regulatory genes*. Some of the most widely conserved associations between transcriptional regulators and specific downstream target genes involve tissue-specific activators and structural genes characteristic of those tissues (Gerhart and Kirschner 1997; Davidson 2001). Although the presence of these associations in distantly related taxa suggests very long-term ($>$ 0.5 billion year) conservation, denser phylogenetic sampling is needed to distinguish this possibility from independent recruitment of downstream genes. (5) *Transcription factor ''switching'' should be a common basis for altered transcription profiles*. Because many transcription factors have overlapping binding specificities (table 4), some point mutations will shift the equilibrium in favor of binding by a different protein. Several cases of such ''transcription factor switching'' have been identified as polymorphisms within human populations (Rockman and Wray 2002). (6) *Negative and quantitative changes in expression should be ''easier'' to achieve than activation of novel expression domains*. Because there are many ways to repress transcription and relatively few ways to activate it, a random change in promoter sequence is more likely to modulate or abolish an existing phase of gene expression than it is to activate a new phase of expression. (7) *The organization of gene networks is robust to perturbations*. Simulations suggest that gene networks are organized in such a way that they produce consistent transcriptional outputs across a range of transcription factor concentrations and transcription factor binding site interactions (von Dassow et al. 2000). If this robustness proves to be a general feature of real gene networks, it may be the result of natural selection to canalize or stabilize transcription against environmental variation and genetic background.

6.10 Promoters Display Complex Macroevolutionary Properties

Reconstructing the macroevolutionary history of promoter structure and function represents an outstanding challenge for studies of developmental evolution (see section 7.2). These changes probably lie at the heart of many important anatomical transformations and innovations (Raff 1986; Carroll, Grenier, and Weatherbee 2001; Davidson 2001; Wilkins 2002). (1) *Promoters are built from a mixture of small-scale mutations and rearrangements*. Populations harbor abundant small-scale variation within promoter regions (nucleotide substitutions, indels, and tandem repeat variants) (see section 2.3). These common forms of mutation can alter transcription, with consequences for fitness (e.g., *FY*: Hamblin and Di Rienzo 2000; *CCR5*: Bamshad et al. 2002; *P450*: Daborn et al. 2002). This ordinary, small-scale variation is likely to be the primary contributor to interspecific differences in promoter sequences. Larger-scale mutations (transposition and chromosomal rearrangement) are less commonly found within populations but can also alter transcription, leading to fitness consequences (e.g., *hsp70*: Lerman et al. 2003). (2) *The binding sites that exist within a single promoter often have different times of origin*. Comparisons between species suggest that promoters often evolve by binding site accretion and turnover (e.g., Ludwig and Kreitman 1995; Rockman and Wray 2002). Importation of intact regulatory modules by transposition or recombination is probably rarer. (3) *Genomic rearrangements can lead to novel expression profiles*. Gene duplications may lead to functional divergence not just of the encoded protein but also of *cis*-regulatory sequences (Ferris and Whitt 1979; Paigen 1989; Force et al. 1999), with several cases now well documented (e.g., *gooseberry* and paralogs: Li and Noll 1994; *Hox3* duplication within Diptera: Stauber, Prell, and Schmidt-Ott 2002). The duplication-degeneration-complementation (DDC) model of promoter evolution (Force et al. 1999) proposes that selection can maintain functionally redundant coding sequences after gene duplication if each copy loses a different promoter module due to random mutation. Although gene families can expand through large-scale processes that duplicate entire promoter regions (polyploidization, chromosomal nondisjunction, and large translocations), local inversions and tandem duplications are more common events (Seoighe et al. 2000). These relatively small genomic rearrangements will often omit some of the *cis*-regulatory sequences surrounding a gene, producing truncated or hybrid promoters at their inception (e.g., *nNOS*: Korneev and O'Shea 2002) and considerably expanding the range of functional outcomes following gene duplication. (4) *The modular structure of promoters facilitates modular changes in expression*. In principle, modular promoter function should allow selection to operate on a discrete aspect of transcription with minimal impact on other aspects of the total transcription profile. Modular promoters may therefore be more evolvable (Gerhart and Kirschner 1997; Stern 2000; Carroll, Grenier, and Weatherbee 2001). (5) *Regulatory recruitment is an important mechanism underlying the evolution of novelty*. Deciphering the genetic basis for anatomical novelty presents a significant challenge in evolutionary biology (Müller and Wagner 1991). One possibility is that new structures require the origin of new genes; the other extreme is that new structures are built entirely by reorganizing the activities or interactions among existing genes. The improbability of de novo origins of functional genes, combined with the observation that evolution generally operates by incremental tinkering (Darwin 1859; Jacob 1977), suggests that the latter process predominates. Several authors have proposed that regulatory changes in development are a crucial and perhaps ubiquitous component in the origin of evolutionary innovations (Britten and Davidson 1971; Duboule and Wilkins 1998; Carroll, Grenier, and Weatherbee 2001; Wilkins 2002). This ''new roles for old genes'' hypothesis (Wray and Lowe 2000) proposes that existing regulatory proteins, including transcription factors, are recruited to build novel features. The requisite genetic variation, namely gains and losses of individual transcription factor binding sites and changes in interactions among proteins bound to these sites, is common within populations (see section 2.3).

7 Future Directions

Despite the challenges unique to studying promoter evolution (see section 5), considerable progress is being made on several fronts. Our ability to study the evolution of transcriptional regulation has increased enormously during the past few years. Biochemical characterizations and expression assays are increasingly feasible in nonmodel organisms, allowing evolutionary comparisons to move beyond sequence inspection to highly informative functional tests. In addition, the number of noncoding sequences available for comparison is increasing exponentially. Whole genome assemblies from related species are proving enormously useful, providing many orthologous intergenic regions for comparison. Importantly, many current areas of ignorance about the evolution of transcriptional regulation are the result of neglect rather than technical limitations. Below we list several important, but poorly understood issues that can be addressed through methods that are practical today.

7.1 Intraspecific Variation

Considerable effort has gone into characterizing general patterns of intraspecific variation in exon and intron sequences and into understanding the mechanisms that shape this variation (Gillespie 1991; Li 1997). Our understanding of variation in these partitions of the genome is now highly quantitative, precise, and, in some cases, predictive. In contrast, most parameters of intraspecific variation have been measured from just one or two promoters, and several basic parameters have never been estimated. An important goal for the near term is to characterize intraspecific variation in promoter sequences: (1) the nature, level, and scope of variation and how it compares to that in other partitions of the genome such as coding sequences, introns, UTRs, and nonregulatory intergenic regions; (2) how much and what components of this variation influence proximate phenotype (transcrip-

tion profile), organismal phenotype (anatomical, physiological, behavioral, etc.), and fitness (fig. 8); (3) the frequencies and heterozygosities of functionally silent versus transcriptionally penetrant and of neutral versus non-neutral allelic variation in promoters; and (4) the relative contributions of mutation, drift, and selection of various kinds to standing variation in promoter sequences. At present, by far the most information on population variation in promoter sequences is available for humans (Cooper 1999; Rockman and Wray 2002). Comparable information is needed from other organisms, not only to determine general patterns of variation but also to enable follow-up studies on functional and selective consequences that cannot be carried out in humans for ethical and practical reasons.

7.2 Reconstructing History

Few studies have analyzed evolutionary changes in promoter structure or function in detail. The frequencies and patterns of evolutionary change in several features are of interest: (1) the spatial distribution for different kinds of binding sites (activator, repressor, architectural factor, general transcription factor), for binding sites of high and low affinity, and among different kinds of modules (activator, repressor, booster, insulator, tetherer, integrator); (2) the composition of binding sites within a promoter (gains, losses, and replacement of individual binding sites), binding site ''switches'' to higher affinity for a different transcription factor, and correlations of such changes with the number of instances of a binding site within a given module or promoter; (3) promoter organization, including changes in spacing and orientation of modules relative to each other and to the basal promoter, the frequency of gains and losses of entire modules, and possible relationships between module turnover and function (activator, repressor, etc.); (4) changes in proximate promoter function including level, timing, and location of expression, as well as gains and losses of entire phases of expression; (5) changes in such modes of transcriptional regulation as constitutive, metabolically inducible, stress inducible, and sexually dimorphic, among others; and (6) changes in the array of upstream regulators that interact with a promoter. Such analyses will be most informative when based on dense taxon sampling and carried out within the context of a robust phylogenetic framework. No promoter has been subjected to a detailed analysis of this kind.

7.3 Structure-Function Relationships

The evolutionary relationship between genetic and phenotypic changes in transcription remains poorly understood, despite a handful of cases where intriguing, but limited, information has been uncovered. Several issues need to be addressed: (1) the proportions of variation (or interspecific differences) in transcription that are heritable, stochastic, and environmentally determined; (2) the extent of maternal influences on gene expression, particularly in embryos; (3) what kinds of mutations give rise to various differences in transcription profiles, such as altered level, altered timing, new location, and environmental sensitivity; (4) the proportion of genetically based variation in transcription that resides in *cis* (within promoter sequences) and *trans* (in the expression or function of regulators that bind to those sequences); (5) the heritability of particular components of transcription, including location, time of onset and cessation, level, and response to inducers or environmental conditions; (6) the degree to which coordinately expressed loci are genetically correlated, for instance through common upstream regulators or the same signal transduction system.

7.4 Relationship to Organismal Phenotype

Despite assertions that mutations within promoter regions constitute the most ''relevant'' (Stern 2000) or ''important'' (Carroll 2000) source of genetic variation, the fraction of phenotypic changes due to mutations within regulatory versus coding sequences is not known, even to a very rough approximation. Estimating this ratio represents an important challenge in molecular evolution. Ultimately, we would like to gauge the broad impact of transcriptional regulation on the evolution of organismal function. Several issues stand out: (1) determining what kinds of phenotypic consequences result from mutations in promoters versus coding sequences, and how these relate to functional classes of encoded proteins (enzyme, transcription factor, ion channel, etc.); (2) conversely, knowing what kinds of mutations in promoters contribute to organismal phenotypes of various kinds (local mutation, chromosomal rearrangement, and transposition, and various classes within each); (3) evaluating how hard it is to achieve a change in a gene expression profile, in particular the number and kind of mutations it takes to shift features such as the timing, level, location, or environmental sensitivity of transcription or to establish a novel phase of transcription; and (4) establishing what fraction of organismal phenotypic changes are due to mutations within regulatory versus coding sequences.

8 Conclusions

Promoter sequences represent for evolutionary biologists a vast and largely uncharted territory within the genome. First principles and a growing body of empirical evidence point squarely toward the evolutionary importance of these regulatory sequences. Because transcriptional regulation is complex, indirect, idiosyncratic, and context dependent, understanding the evolutionary mechanisms that shape promoter sequences will require a thorough appreciation of molecular mechanisms as well as the use of comparative data from promoter sequences, biochemical assays, and functional tests. The conceptual and empirical challenges to studying promoter evolution are significant, but well worth tackling. The insights into evolutionary history and mechanisms that will emerge from detailed analyses of promoter evolution are potentially enormous. This information will be essential for a complete understanding of the evolution of the genotype-phenotype relationship. Changes in promoter function are likely to be an important component of reproductive isolation, the

evolution of morphology and physiology, the origin of phenotypic plasticity, and the genetic basis of evolutionary novelties. Yet to date nearly everything we know about the evolution of promoters has come from biologists with relatively little background or interest in evolutionary biology. It is time for the molecular evolution community to seize the opportunities that promoters offer to expand our understanding and appreciation of the evolution of genomes and organisms.

Acknowledgments

Lynn Angerer, Ann Rouse, David Des Marais, Daven Presgraves, Tania Rehse, and Jay Storz provided perceptive comments and helped locate references. Two anonymous reviewers offered constructive advice. Research in G.A.W.'s lab is supported by the National Science Foundation and by the National Aeronautics and Space Administration.

Literature Cited

Abouheif, E. H., and G. A. Wray. 2002. The developmental genetic basis for the evolution of wing polyphenism in ants. Science **297**:249–252.

Abraham, I., and W. W. Doane. 1978. Genetic regulation of tissue-specific expression of amylase structural genes in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **75**:4446–4450.

Abu-Shaar, M., H. D. Ryoo, and R. S. Mann. 1999. Control of the nuclear localization of extradenticle by competing nuclear import and export signals. Genes Dev. **13**:935–945.

Abzhanov, A., and T. C. Kaufman. 2000. Crustacean (malacostracan) *Hox* genes and the evolution of the arthropod trunk. Development **127**:2239–2248.

Akashi, H. 2001. Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11**:660–666.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. 2002. The molecular biology of the cell. Garland Publishing, New York.

Allendorf, F. W., K. L. Knudsen, and R. F. Leary. 1983. Adaptive significance of differences in the tissue-specific expression of a phosphoglucomutase gene in rainbow trout. Proc. Natl. Acad. Sci. USA **80**:1397–1400.

Allendorf, F. W., K. L. Knudsen, and S. R. Phelps. 1982. Identification of a gene regulating the tissue expression of a phosphoglucomutase locus in rainbow trout. Genetics **102**:259–268.

Andersson, C. R., E. O. Jensen, D. J. Llewellyn, E. S. Dennis, and W. J. Peacock. 1996. A new hemoglobin gene from soybean: a role for hemoglobin in all plants. Proc. Natl. Acad. Sci. USA **93**:5682–5687.

Angerer, L. M., D. W. Oleksyn, A. M. Levine, X. Li, W. H. Klein, and R. C. Angerer. 2001. Sea urchin goosecoid function links fate specification along the animal-vegetal and oral-aboral embryonic axes. Development **128**:4393–4404.

Anisimov, S. V., M. V. Volkova, L. V. Lenskaya, V. K. Khavinson, D. V. Solovieva, and E. I. Schwartz. 2001. Age-associated accumulation of the apolipoprotein C-III gene T-455C polymorphism C allele in a Russian population. J. Gerontol. A. Biol. Sci. Med. Sci. **56**:B27–B32.

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf, and S. Brenner. 1995. Detecting conserved regulatory elements with the model genome of the Japanese pufferfish, *Fugu rubipes*. Proc. Natl. Acad. Sci. USA **92**:1684–1688.

Arbeitman, M. N., E. E. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. 2002. Gene expression during the life cycle of *Drosophila melanogaster*. Science **297**:2270–2275.

Arnone, M. I., and E. H. Davidson. 1997. The hardwiring of development: organization and function of genomic regulatory systems. Development **124**:1851–1864.

Arnone, M. I., E. L. Martin, and E. H. Davidson. 1998. *Cis*-regulation downstream of cell type specification: a single compact element controls the complex expression of the *CyIIa* gene in sea urchin embryos. Development **125**:1381–1395.

Atchison, M. L. 1988. Enhancers: mechanisms of action and cell specificity. Annu. Rev. Cell Biol. **4**:127–153.

Averof, M., and N. H. Patel. 1997. Crustacean appendage evolution associated with changes in *Hox* gene expression. Nature **388**:682–686.

Avila, S., M. C. Casero, R. Fernandez-Canton, and L. Sastre. 2002. Transactivation domains are not functionally conserved between vertebrate and invertebrate serum response factors. Eur. J. Biochem. **269**:3669–3677.

Babich, V., N. Aksenov, V. Alexeenko, S. L. Oei, G. Buchlow, and N. Tomilin. 1999. Association of some potential hormone response elements in human genes with the Alu family repeats. Gene **239**:341–349.

Bamshad, M. J., S. Mummidi, E. Gonzalez et al. (11 co-authors). 2002. A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. Proc. Natl. Acad. Sci. USA **99**:10539–10544.

Barrier, M., R. H. Robichaux, and M. D. Purugganan. 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proc. Natl. Acad. Sci. USA **98**:10208–10213.

Beckers, J., and D. Duboule. 1998. Genetic analysis of a conserved sequence in the *HoxD* complex: regulatory redundancy or limitations of the transgenic approach? Dev. Dyn. **213**:1–11.

Beldade, P., P. M. Brakefield, and A. D. Long. 2002. Contribution of *Distal-less* to quantitative variation in butterfly eyespots. Nature **415**:315–318.

Bell, A. C., and G. Felsenfeld. 1999. Stopped at the border: boundaries and insulators. Curr. Opin. Genet. Dev. **9**:191–198.

Bell, S. D., and S. P. Jackson. 1998. Transcription in Archaea. Cold Spring Harbor Symp. Quant. Biol. **63**:41–51.

Belting, H.-G., C. S. Shashikant, and F. H. Ruddle. 1998. Modification of expression and *cis*-regulation of *Hoxc8* in the evolution of diverged axial morphology. Proc. Natl. Acad. Sci. USA **95**:2355–2360.

Benbrook, D. M., and N. C. Jones. 1990. Heterodimer formation between CREB and JUN proteins. Oncogene **5**:295–302.

Bender, W., M. Akam, F. Karch, P. A. Beachy, M. Peifer, P. Spierer, E. B. Lewis, and D. S. Hogness. 1983. Molecular-genetics of the bithorax complex in *Drosophila melanogaster*. Science **221**:23–29.

Benecke, A., C. Gaudon, and H. Gronemeyer. 2001. Transcriptional integration of hormone and metabolic signals by nuclear receptors. Pp. 167–214 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.

Benezra, R., R. L. Davis, D. Lockshon, D. L. Turner, and H. Weintraub. 1990. The protein Id: a negative regulator of helix-loop-helix DNA-binding proteins. Cell **61**:49–59.

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.

Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celnick, M. Levine, G. M. Rubin, and M. B. Eisen. 2002.

Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA **99**: 757–767.

Bernstein, B. E., J. K. Tong, and S. L. Schreiber. 2000. Genome-wide studies of histone deacetylase function in yeast. Proc. Natl. Acad. Sci. USA **97**:13708–13713.

Berthelsen, J., V. Zappavigna, E. Feretti, F. Mavilio, and F. Blasi. 1998. The novel homeoprotein Prep1 modulates Pbx-Hox protein cooperativity. EMBO J. **17**:1434–1445.

Betran, E., K. Thornton, and M. Long. 2002. Retroposed new genes out of the X in Drosophila. Genome Res. **12**:1854–1859.

Bharathan, G., T. E. Goliber, C. Moore, S. Kessler, T. Pham, and N. R. Sinha. 2002. Homologies in leaf form inferred from *KNOXI* gene expression during development. Science **296**: 1858–1860.

Bharathan, G., B.-J. Janssen, E. A. Kellogg, and N. Sinha. 1997. Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Proc. Natl. Acad. Sci. USA **94**:13749–13753.

Biggin, M. D., and W. McGinnis. 1997. Regulation of segmentation and segmental identity by Drosophila homeoproteins: the role of DNA binding in functional activity and specificity. Development **124**:4425–4433.

Birnbaum, K., P. N. Benfey, and D. E. Shasha. 2001. *cis* element/transcription factor analysis (*cis*/TF): a method for discovering transcription factor/*cis* element relationships. Genome Res. **11**:1567–1573.

Blumenthal, T. 1998. Gene clusters and polycistronic transcription in eukaryotes. Bioessays **20**:480–487.

Bonifer, C. 2000. Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required? Trends Genet. **16**:310–315.

Brakefield, P. M., J. Gates, D. Keys, F. Kesbeke, P. J. Wijngaarden, A. Monteiro, V. French, and S. B. Carroll. 1996. Development, plasticity and evolution of butterfly eyespot patterns. Nature **384**:236–242.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science **296**:752–755.

Brickman, J. M., M. Clements, R. Tyrell, D. McNay, K. Woods, J. Warner, A. Stewart, R. S. P. Beddington, and M. Dattani. 2001. Molecular effects of novel mutations in Hesx1/HESX1 associated with human pituitary disorders. Development **128**: 5189–5199.

Britten, R. J. 1997. Mobile elements inserted in the distant past have taken on important functions. Gene **205**:177–182.

Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. Science **165**:349–357.

———. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. **46**:111–138.

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene **238**:115–134.

Brunetti, C. R., J. E. Selegue, A. Monteiro, V. French, P. M. Brakefield, and S. B. Carroll. 2001. The generation and diversification of butterfly eyespot color patterns. Curr. Biol. **11**:1578–1585.

Buckwold, V. E., Z. C. Xu, T. S. B. Yen, and J. H. Ou. 1997. Effects of a frequent double-nucleotide basal core promoter mutation and its putative single-nucleotide precursor mutations on hepatitis B virus gene expression and replication. J. Gen. Virol. **78**:2055–2065.

Budd, G. E. 1999. Does evolution in body patterning genes drive morphological change—or vice versa? Bioessays **21**: 326–332.

Buggs, C., N. Nasrin, A. Mode, P. Tollet, H.-F. Zhao, J.-Å. Gustafsson, and M. Alexander-Bridges. 1998. IRE-ABP (insulin response element-A binding protein), An SRY-like protein, inhibits C/EBPα (CCAAT/enhancer-binding protein α)-stimulated expression of the sex-specific cytochrome P450 2C12 gene. Mol. Endocrinol. **12**:1294–1309.

Bürglin, T. 1997. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iriquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic Acids Res. **25**:4173–4180.

Burke, A. C., C. E. Nelson, B. A. Morgan, and C. Tabin. 1995. *Hox* genes and the evolution of vertebrate axial morphology. Development **121**:333–346.

Burstin, J., D. De Vienne, P. Dubreuil, and C. Damerval. 1994. Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines. Theor. Appl. Genet. **89**:943–950.

Bush, R. M., and K. Paigen. 1992. Evolution of β-*glucuronidase* regulation in the genus *Mus*. Evolution **46**:1–15.

Calhoun, V. C., A. Stathopoulos, and M. Levine. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proc. Natl. Acad. Sci. USA **99**:9243–9247.

Carey, M., and S. T. Smale. 2000. Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Carrión, A. M., W. A. Link, F. Ledo, B. Mellström, and J. R. Naranjo. 1999. DREAM is a Ca^{2+}-regulated transcriptional repressor. Nature **398**:80–84.

Carroll, S. B. 1995. Homeotic genes and the evolution of arthropods and chordates. Nature **376**:479–485.

———. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell **101**:577–580.

Carroll, S. B., J. K. Grenier, and S. D. Weatherbee. 2001. From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Science, Malden, Mass.

Caspi, A., J. McClay, T. E. Moffitt, J. Mill, J. Martin, I. W. Craig, A. Taylor, and R. Poulton. 2002. Role of genotype in the cycle of violence in maltreated children. Science **297**:851–854.

Cavalieri, D., J. P. Townsend, and D. L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. Proc. Natl. Acad. Sci. USA **97**:12369–12374.

Cavener, D. R. 1992. Transgenic animal studies on the evolution of genetic regulatory circuitries. Bioessays **14**:237–244.

Cereb, N., and S. Y. Yang. 1994. The regulatory complex of HLA class I promoters exhibits locus-specific conservation with limited allelic variation. J. Immunol. **152**:3873–3883.

Chen, G., and A. J. Courey. 2000. Groucho/TLE family proteins and transcriptional repression. Gene **249**:1–16.

Chiu, C. H., H. Schneider, J. L. Slightom, D. L. Gumucio, and M. Goodman. 1997. Dynamics of regulatory evolution in primate *beta-globin* gene clusters: *cis*-mediated acquisition of simian gamma fetal expression patterns. Gene **205**:47–57.

Choo, Y., and A. Klug. 1997. Physical basis of a protein-DNA recognition code. Curr. Opin. Struct. Biol. **7**:117–125.

Christophides, G. K., I. Livadras, C. Savakis, and K. Komitopolou. 2000. Two medfly promoters that have originated by recent gene duplications drive distinct sex, tissue and temporal expression patterns. Genetics **156**:173–182.

Chung, Y. D., H. C. Kwon, K. W. Chung, S. J. Kim, K. J. Kim, and C. C. Lee. 1996. Identification of ovarian enhancer-binding factors which bind to ovarian enhancer 1 of the Drosophila genes *yp1* and *yp2*. Mol. Gen. Genet. **251**:347–351.

Clark, A. G. 1990. Genetic components of variation in energy storage in *Drosophila melanogaster*. Evolution **44**:637–650.

Coller, H. A., C. Grandori, P. Tamayo, T. Colbert, E. S. Lander,

R. N. Eisenman, and T. R. Golub. 2000. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc. Natl. Acad. Sci. USA **97**:3260–3265.

Conlon, F. L., L. Fairclough, B. M. J. Price, E. S. Casey, and J. C. Smith. 2001. Determinants of T box protein specificity. Development **128**:3749–3758.

Cooper, D. N. 1999. Human gene evolution. Academic Press, San Diego, Calif.

Core, N., B. Charroux, A. McCormick, C. Vola, L. Fasano, M. P. Scott, and S. Kerridge. 1997. Transcriptional regulation of the Drosophila homeotic gene *teashirt* by the homeodomain protein Fushi tarazu. Mech. Dev. **68**:157–172.

Costa, P., and C. Plomion. 1999. Genetic analysis of needle proteins in maritime pine. 2. Variation in protein accumulation. Silvae Genet. **48**:146–150.

Courey, A. J. 2001. Regulatory transcription factors and *cis*-regulatory regions. Pp. 17–34 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.

Cowell, L. G., T. B. Kepler, M. Janitz, R. Lauster, and N. A. Mitchison. 1998. The distribution of variation in regulatory gene segments, as present in MHC class II promoters. Genome Res. **8**:124–134.

Cowen, L. E., D. Sanglard, D. Calabrese, C. Sirjusingh, J. B. Anderson, and L. M. Kohn. 2000. Evolution of drug resistance in experimental populations of *Candida albicans*. J. Bacteriol. **182**:1515–1522.

Cowles, C. R., J. N. Hirshhorn, D. Altshuler, and E. S. Lander. 2002. Detection of regulatory variation in mouse genes. Nat. Genet. **32**:432–437.

Crawford, D. L., J. A. Segal, and J. L. Barnett. 1999. Evolutionary analysis of TATA-less proximal promoter function. Mol. Biol. Evol. **16**:194–207.

Czerny, T., G. Schaffner, and M. Busslinger. 1993. DNA sequence recognition by pax proteins: bipartite structure of the paired domain and its binding site. Genes Dev. **7**:2048–2061.

Daborn, P. J., J. L. Yen, M. R. Bogwitz, G. L. Goff, E. Feil, S. Jeffers, N. Tijet, T. Perry, D. Heckel, P. Batterham, R. Feyereisen, T. G. Wilson, and R. H. French-Constant. 2002. A single P450 allele associated with insecticide resistance in Drosophila. Science **297**:2253–2225.

Dailey, L., and C. Basilico. 2001. Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. J. Cell. Physiol. **186**:315–328.

Damerval, C., A. Maurice, J. M. Josse, and D. De Vienne. 1994. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics **137**:289–301.

Darwin, C. 1859. On the origin of species by means of natural selection. John Murray, London.

Davidson, E. H. 2001. Genomic regulatory systems: development and evolution. Academic Press, San Diego, Calif.

Dawes, R., I. Dawson, F. Falciani, G. Tear, and M. Akam. 1994. *Dax*, a locust *hox* gene related to *fushi-tarazu* but showing no pair-rule expression. Development **120**:1561–1572.

Dawson, S. J., P. J. Morris, and D. S. Latchman. 1996. A single amino acid change converts a repressor into an activator. J. Biol. Chem. **271**:11631–11633.

De Vienne, D., B. Bost, J. Fievet, M. Zivy, and C. Dillmann. 2001. Genetic variability of proteome expression and metabolic control. Plant Physiol. Biochem. **39**:271–283.

D'Elia, A. V., G. Tell, I. Paron, L. Pellizzari, R. Lonigro, and G. Damante. 2002. Missense mutations of human homeoboxes: a review. Hum. Mutat. **18**:361–374.

Denizot, Y., E. Pinaud, C. Aupetit, C. Le Morvan, E. Magnoux, J. C. Aldigier, and M. Cogne. 2001. Polymorphism of the human alpha1 immunoglobulin gene 3′ enhancer hs1,2 and its relation to gene expression. Immunology **103**:35–40.

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.

DeRobertis, E. M., and Y. Sasai. 1996. A common plan for dorsoventral patterning in Bilateria. Nature **380**:37–40.

Di Gregorio, A., J. C. Corbo, and M. Levine. 2001. The regulation of *forkhead/HNf-3 beta* expression in the *Ciona* embryo. Dev. Biol. **229**:31–43.

Dickinson, W. J. 1988. On the architecture of regulatory systems: evolutionary insights and implications. Bioessays **8**:204–208.

DiLeone, R. J., L. B. Russell, and D. M. Kingsley. 1998. An extensive 3′ regulatory region controls expression of *Bmp5* in specific anatomical structures of the mouse embryo. Genetics **148**:401–408.

Dillon, N., and P. Sabbattini. 2000. Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. Bioessays **22**:657–665.

Dobzhansky, T. 1936. Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics **21**:113–135.

Doebley, J., and L. Lukens. 1998. Transcriptional regulators and the evolution of plant form. Plant Cell **10**:1075–1082.

Droge, P., and B. Muller-Hill. 2001. High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. Bioessays **23**:179–183.

Duboule, D. 1994. Guidebook to the homeobox genes. Oxford University Press, Oxford.

Duboule, D., and A. S. Wilkins. 1998. The evolution of 'bricolage.' Trends Genet. **14**:54–59.

Dudareva, N., L. Cseke, V. M. Blanc, and E. Pichersky. 1996. Evolution of floral scent in ?? C*larkia*: novel patterns of S-linalool synthase gene expression in the *C. breweri* flower. Plant Cell **8**:1137–1148.

Dynan, W. S. 1989. Modularity in promoters and enhancers. Cell **58**:1–4.

Enard, W., P. Khaitovich, J. Klose et al. (13 co-authors). 2002*a*. Intra- and interspecific variation in primate gene expression patterns. Science **296**:340–343.

Enard, W., M. Przeworski, S. E. Fisher, C. S. L. Lai, V. Wiebe, T. Kitano, A. P. Monaco, and S. Pääbo. 2002*b*. Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418**:869–872.

Fairall, L., and J. W. R. Schwabe. 2001. DNA binding by transcription factors. Pp. 65–84 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.

Falciani, F., B. Hausdorf, R. Schröder, M. Akam, D. Tautz, R. Denell, and S. Brown. 1996. Class 3 *Hox* genes in insects and the origin of *zen*. Proc. Natl. Acad. Sci. USA **93**:8479–8484.

Fang, H., and B. P. Brandhorst. 1996. Expression of the actin gene family in embryos of the sea urchin *Lytechinus pictus*. Dev. Biol. **173**:306–317.

Fang, S., A. Takahashi, and C.-I. Wu. 2002. A mutation in the promoter of *desaturase 2* is correlated with sexual isolation between Drosophila behavioral races. Genetics **162**:781–784.

Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. USA **96**:9721–9726.

Ferkowicz, M. J., and R. Raff. 2001. *Wnt* gene expression in sea urchin development: heterochronies associated with the evolution of developmental mode. Evol. Dev. **3**:24–33.

Ferrigno, O., T. Virolle, Z. Djabari, J. P. Ortonne, R. J. White, and D. Aberdam. 2001. Transposable B2 SINE elements can

provide mobile RNA polymerase II promoters. Nat. Genet. **28**:77–81.
Ferris, S. D., and G. S. Whitt. 1979. Evolution of the differential regulation of duplicate genes following polyploidization. J. Mol. Evol. **12**:267–317.
Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.
Fitch, D. H. A. 1997. Evolution of male tail development in rhabditid nematodes related to *Caenorhabditis elegans*. Syst. Biol. **46**:145–179.
Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.
Flores-Saaib, R. D., S. Jia, and A. J. Courey. 2001. Activation and repression by the C-terminal domain of Dorsal. Development **128**:1869–1879.
Force, A., M. Lynch, F. B. Pickett, A. Amores, Y.-L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**:1531–1545.
Foulkes, N. S., and P. Sassone-Corsi. 1992. More is better: activators and repressors from the same gene. Cell **68**:411–414.
Frasch, M., X. Chen, and T. Lufkin. 1995. Evolutionary-conserved enhancers direct region-specific expression of the murine *Hoxa-1* and *Hoxa-2* loci in both mice and Drosophila. Development **121**:957–974.
Frazer, K. A., J. B. Sheehan, R. P. Stokowski, X. Chen, R. Hosseini, J. F. Cheng, S. P. Fodor, D. R. Cox, and N. Patil. 2001. Evolutionarily conserved sequences on human chromosome 21. Genome Res. **11**:1651–1659.
Fry, C. J., and P. J. Farnham. 1999. Context-dependent transcriptional regulation. J. Biol. Chem. **274**:29583–29586.
Fu, Y.-X., and W.-H. Li. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693–709.
Fullerton, S. M., A. Bartoszewicz, G. Ybazeta, Y. Horikawa, G. I. Bell, K. K. Kidd, N. J. Cox, R. R. Hudson, and A. Di Rienzo. 2002. Geographic and haplotype structure of candidate type 2 diabetes-susceptibility variants at the *calpain-10* locus. Am. J. Hum. Genet. **70**:1096–1106.
Galant, R., and S. B. Carroll. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. Nature **415**:910–913.
Gehring, W. J., and K. Ikeo. 1999. Pax6: mastering eye morphogenesis and eye evolution. Trends Genet. **15**:371–377.
Gerard, M., J. Zakany, and D. Duboule. 1997. Interspecies exchange of a Hoxd enhancer in vivo induces premature transcription and anterior shift of the sacrum. Dev. Biol. **190**:32–40.
Gerber, S., F. Fabre, and C. Planchon. 2000. Genetics of seed quality in soybean analysed by capillary gel electrophoresis. Plant Sci. **152**:181–189.
Gerhart, J., and M. Kirschner. 1997. Cells, embryos, and evolution: toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Blackwell Science, Malden, Mass.
Gibson, G. 1996. Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor. Popul. Biol. **49**:58–89.
Gilbert, S. F. 2000. Developmental biology. Sinauer Associates, Sunderland, Mass.
———. 2001. Ecological developmental biology: developmental biology meets the real world. Dev. Biol. **233**:1–12.
Gill, G., and M. Ptashne. 1987. Mutants of GAL4 protein altered in an activation function. Cell **51**:121–126.
Gillespie, J. H. 1991. The causes of molecular evolution. Oxford University Press, New York.
Giordano, M., C. Marchetti, E. Chiorboli, G. Bona, and P. Momigliano Richiardi. 1997. Evidence for gene conversion in the generation of extensive polymorphism in the promoter of the growth hormone gene. Hum. Genet. **100**:249–255.
Glass, C. K., D. W. Rose, and M. G. Rosenfeld. 1997. Nuclear receptor coactivators. Curr. Opin. Cell Biol. **9**:222–232.
Gonzalez, P., P. V. Rao, S. B. Nunez, and J. S. Zigler, Jr. 1995. Evidence for independent recruitment of zeta-crystallin/quinone reductase (CRYZ) as a crystallin in camelids and hystricomorph rodents. Mol. Biol. Evol. **12**:773–781.
Goodyer, C. G., G. Zogopolos, G. Schwartzbauer, H. Zheng, G. N. Hendy, and R. K. Menon. 2001. Organization and evolution of the human growth hormone receptor 5′-flanking region. Endocrinology **142**:1923–1934.
Grandori, C., S. M. Cowley, L. P. James, and R. N. Eisenman. 2000. The Myc/Max/Mad network and the transcriptional control of cell behavior. Annu. Rev. Cell. Dev. Biol. **16**:653–699.
Gray, S., and M. Levine. 1996. Transcriptional repression in development. Curr. Opin. Cell Biol. **8**:358–364.
Grbic, M., L. M. Nagy, and M. R. Strand. 1998. Polyembryonic insect development: insect pattern formation in a cellularised environment. Dev. Genes Evol. **208**:69–81.
Grosveld, F., M. Antoniou, M. Berry et al. (16 co-authors). 1993. The regulation of human globin gene switching. Phil. Trans. R. Soc. Lond. Ser. B **339**:183–191.
Gstaiger, M., L. Knoepfl, O. Georgiev, W. Schaffner, and C. M. Hovens. 1995. A B-cell co-activator of octamer-binding transcription factors. Nature **373**:360–362.
Gu, W., and R. G. Roeder. 1997. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. Cell **90**:595–606.
Gu, Z., D. Nicolae, H. H.-S. Lu, and W.-H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **18**:609–613.
Guardiola, J., A. Maffei, R. Lauster, N. A. Mitchison, R. S. Accolla, and S. Sartoris. 1996. Functional significance of polymorphism among MHC class II gene promoters. Tissue Antigens **48**:615–625.
Hahn, M. W., M. D. Rausher, and C. W. Cunningham. 2002. Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. Genetics **161**:11–20.
Hahn, M. W., J. E. Stajich, and G. A. Wray. 2003. The effects of selection against spurious transcription factor binding sites. Mol. Biol. Evol. (in press).
Hamblin, M. T., and A. DiRienzo. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66**:1669–1679.
Hancock, J., P. Shaw, F. Benneton, and G. Dover. 1999. High sequence turnover in the regulatory regions of the developmental gene *hunchback* in insects. Mol. Biol. Evol. **16**:253–265.
Hanes, S. D., G. Riddihough, D. Ish-Horowicz, and R. Brent. 1994. Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. Mol. Cell. Biol. **14**:3364–3375.
Hanna-Rose, W., and U. Hansen. 1996. Active repression mechanisms of eukaryotic transcription factors. Trends Genet. **12**:229–234.
Hardison, R. C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. **16**:369–372.
Hariri, A. R., V. S. Mattay, A. Tessitore, B. Kolachana, F. Fera, D. Goldman, M. F. Egan, and D. R. Weinberger. 2002. Serotonin transporter genetic variation and the response of the human amygdala. Science **297**:400–403.

Harrison, S. C. 1991. A structural taxonomy of DNA-binding domains. Nature **353**:715–719.
Haudek, S. B., B. E. Natmessnig, H. Redl, G. Schlag, and B. P. Giroir. 1998. Genetic sequences and transcriptional regulation of the TNFA promoter: comparison of human and baboon. Immunogenetics **48**:202–207.
Hill, T. A., C. D. Day, S. C. Zondlo, A. G. Thackeray, and V. F. Irish. 1998. Discrete spatial and temporal *cis*-acting elements regulate transcription of the *Arabidopsis* floral homeotic gene *APETELA3*. Development **125**:1711–1721.
Hizver, J., H. Rozenberg, F. Frolow, D. Rabinovich, and Z. Shakked. 2001. DNA bending by an adenine-thymine tract and its role in gene regulation. Proc. Natl. Acad. Sci. USA **98**:8490–8495.
Hodgkin, J. 1992. Genetic sex determination mechanisms and evolution. Bioessays **14**:253–261.
Holland, N. D., and L. Z. Holland. 1999. Amphioxus and the utility of molecular genetic data for hypothesizing body part homologies between distantly related animals. Am. Zool. **39**:630–640.
Holstege, F. C. P., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95**:717–728.
Hope, I. A., and K. Struhl. 1986. Functional dissection of a eukaryotic transcriptional activator, GCN4 of yeast. Cell **46**:885–894.
Houchens, C. R., W. Montigny, L. Zeltser, L. Dailey, J. M. Gilbert, and N. H. Heintz. 2000. The dhfr ori beta-binding protein RIP60 contains 15 zinc fingers: DNA binding and looping by the central three fingers and an associated proline-rich region. Nucleic Acids Res. **28**:570–581.
Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.
Hunt, G. M., D. Johnson, and C. T. Tiemesse. 2001. Characterisation of the long terminal repeat regions of South African human immunodeficiency virus type 1 isolates. Virus Genes **23**:27–34.
Indovina, P., F. Megiorni, P. Ferrante, I. Apollonio, F. Petronzelli, and M. C. Mazzilli. 1998. Different binding of NF-Y transcriptional factor to *DQA1* promoter variants. Hum. Immunol. **59**:758–767.
Iyer, V. R., C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature **409**: 533–538.
Jackson, S. P., and R. Tjian. 1988. O-glycosylation of eukaryotic transcription factors: implications for mechanisms of transcriptional regulation. Cell **55**:125–133.
Jackson-Fisher, A. J., C. Chitikila, M. Mitra, and B. F. Pugh. 1999. A role for TBP dimerization in preventing unregulated gene expression. Mol. Cell **3**:717–727.
Jacob, F. 1977. Evolution and tinkering. Science **196**:1161–1166.
Jacob, F., and J. Monod. 1961. On the regulation of gene activity. Cold Spring Harbor Symp. Quant. Biol. **26**:193–211.
Jacobs, J. J., and M. Van Lohuizen. 1999. Cellular memory of transcriptional states by polycomb-group proteins. Semin. Cell. Dev. Biol. **10**:227–235.
James, L., and R. N. Eisenman. 2002. Myc and Mad bHLHZ domains possess identical DNA-binding specificities but only partially overlapping functions in vivo. Proc. Natl. Acad. Sci. USA **99**:10429–10434.
Jareborg, N., E. Birney, and R. Durbin. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9**:815–824.
Jaynes, J. B., and P. H. O'Farrell. 1991. Active repression of transcription by the engrailed homeodomain protein. EMBO J. **10**:1427–1433.
Jeeninga, R. E., M. Hoogenkamp, M. Armand-Ugon, M. De Baar, K. Verhoef, and B. Berkhout. 2000. Functional differences between the long terminal repeat transcriptional promoters of human immunodeficiency virus type 1 subtypes A through G. J. Virol. **74**:3740–3751.
Jenkins, D. L., C. A. Ortori, and J. F. Y. Brookfield. 1995. A test for adaptive change in DNA sequences controlling transcription. Proc.R. Soc. Lond. Ser. B Biol. Sci. **261**:203–207.
Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29**:389–395.
Johnson, N. A., and A. H. Porter. 2000. Rapid speciation via parallel, directional selection on regulatory genetic pathways. J. Theor. Biol. **205**:527–542.
Jones, P. A., and D. Takai. 2001. The role of DNA methylation in mammalian epigenetics. Science **293**:1068–1070.
Jones, S., P. Van Heyningen, H. M. Berman, and J. M. Thornton. 1999. Protein–DNA interactions: a structural analysis. J. Mol. Biol. **287**:877–896.
Jordan, I. K., and J. F. McDonald. 1998. Interelement selection in the regulator region of the *copia* retrotransposon. J. Mol. Evol. **47**:670–676.
Kadosh, D., and K. Struhl. 1998. Targeted recruitment of the Sin3-Rpd3 histone deacetylase complex generates a highly localized domain of repressed chromatin in vivo. Mol. Cell. Biol. **18**:5121–5127.
Kajiya, Y., K. Hamasaki, K. Nakata et al. (11 co-authors). 2001. A long-term follow-up analysis of serial core promoter and precore sequences in Japanese patients chronically infected by hepatitis B virus. Digest. Dis. Sci. **46**:509–515.
Kammandel, B., K. Chowdhury, A. Stoykova, S. Aparicio, S. Brenner, and P. Gruss. 1999. Distinct *cis*-essential modules direct the time-space pattern of the *Pax6* gene activity. Dev. Biol. **205**:79–97.
Karp, C. L., A. Grupe, E. Schadt et al. (13 co-authors). 2000. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat. Immunol. **1**: 221–226.
Kayo, T., D. B. Allison, R. Weindruch, and T. A. Prolla. 2001. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. Proc. Natl. Acad. Sci. USA **98**:5093–5098.
Kazazian, H. H. 1990. The thalassemia syndromes: molecular basis and prenatal diagnosis in 1990. Semin. Hematol. **27**: 209–228.
Keys, D. N., D. L. Lewis, J. E. Selegue, B. J. Pearson, L. V. Goodrich, R. L. Johnson, J. Gates, M. P. Scott, and S. B. Carroll. 1999. Recruitment of a *hedgehog* regulatory circuit in butterfly eyespot evolution. Science **283**:532–534.
Kidwell, M. G., and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. USA **94**:7704–7711.
Kim, J. 2001. Macro-evolution of the hairy enhancer in Drosophila species. J. Exp. Zool. **291**:175–185.
Kim, J., J. Q. Kerr, and G.-S. Min. 2000. Molecular heterochrony in the early development of Drosophila. Proc. Natl. Acad. Sci. USA **97**:212–216.
Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.
King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107–116.
Kirchhamer, C. V., L. D. Bogarad, and E. H. Davidson. 1996. Developmental expression of synthetic cis-regulatory systems

composed of spatial control elements from two different genes. Proc. Natl. Acad. Sci. USA **93**:13849–13854.
Kirchhamer, C. V., C.-H. Yuh, and E. H. Davidson. 1996. Modular *cis*-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc. Natl. Acad. Sci. USA **93**:9322–9328.
Kissinger, J. C., and R. A. Raff. 1998. Evolutionary changes in sites and timing of actin gene expression in embryos of the direct- and indirect-developing sea urchins, *Heliocidaris erythrogramma* and *H. tuberculata*. Dev. Genes Evol. **208**:82–93.
Klarenberg, A. J., K. Sikkema, and W. Scharloo. 1987. Functional significance of regulatory map and structural *amy* variants in *Drosophila melanogaster*. Heredity **58**:383–389.
Klein, W. H., and X. T. Li. 1999. Function and evolution of Otx proteins. Biochem. Biophys. Res. Commun. **258**:229–233.
Klose, J., C. Nock, M. Herrmann et al. (13 co-authors). 2002. Genetic analysis of the mouse brain proteome. Nat. Genet. **30**:385–393.
Kmita, M., T. Kondo, and D. Duboule. 2000. Targeted inversion of a polar silencer within the *HoxD* complex re-allocates domains of enhancer sharing. Nat. Genet. **26**:451–454.
Knoepfler, P. S., and M. P. Kamps. 1995. The pentapeptide motif of hox proteins is required for cooperative DNA binding with pbx1, physically contacts pbx1 and enhances DNA binding by pbx1. Mol. Cell. Biol. **15**:5811–5819.
Kopp, A., I. Duncan, and S. B. Carroll. 2000. Genetic control and evolution of sexually dimorphic characters in Drosophila. Nature **408**:553–559.
Korneev, S., and M. O'Shea. 2002. Evolution of nitric oxide synthase regulatory genes by DNA inversion. Mol. Biol. Evol. **19**:1228–1233.
Kramer, S. G., T. M. Jinks, P. Schedl, and J. P. Gergen. 1999. Direct activation of *Sex-lethal* transcription by the Drosophila runt protein. Development **126**:191–200.
Kuras, L., and K. Struhl. 1999. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. Nature **399**:609–613.
Lander, E. S., L. M. Linton, and B. Birren et al. (242 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.
Latchman, D. S. 1998. Eukaryotic transcription factors. Academic Press, San Diego, Calif.
Laurie-Ahlberg, C. C., and G. C. Bewley. 1983. Naturally occurring genetic variation affecting the expression of *sn-glycerol-3-phosphate dehydrogenase* in *Drosophila melanogaster*. Biochem. Genet. **21**:943–961.
Laurie-Ahlberg, C. C., G. Maroni, G. C. Bewley, J. C. Lucchesi, and B. S. Weir. 1980. Quantitative genetic variation of enzyme activities in natural populations of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **77**:1073–1077.
Lawton-Rauh, A. L., E. R. Alvarez-Buylla, and M. D. Purugganan. 2000. Molecular evolution of flower development. Trends Ecol. Evol. **15**:144–149.
Lee, H., S. N. Cho, H. E. Bang, J. H. Lee, G. H. Bai, S. J. Kim, and J. D. Kim. 2000. Exclusive mutations related to isoniazid and ethionamide resistance among *Mycobacterium tuberculosis* isolates from Korea. Eur. J. Clin. Microbiol. Infect. Dis. **17**:508–511.
Lee, J. E. 1997. Basic helix-loop-helix genes in neural development. Curr. Opin. Neurobiol. **7**:13–20.
Lee, T. I., N. J. Rinaldi, R. Robert et al. (20 co-authors). 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science **298**:799–804.
Lee, T. I., and R. A. Young. 2000. Transcription of eukaryotic protein-coding genes. Annu. Rev. Genet. **34**:77–137.
Lemon, B., and R. Tjian. 2000. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. **14**:2551–2569.
Lerman, D. N., P. Michalak, A. B. Helin, B. R. Bettencourt, and M. E. Feder. 2003. Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. Mol. Biol. Evol. **20**:135–144.
Lescot, M., P. Dehais, G. Thijs, K. Marchal, Y. Moreau, Y. van de Peer, P. Rouze, and S. Rombauts. 2002. PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. **30**:325–327.
Lettice, L. A., T. Horikoshi, S. J. H. Heaney et al. (18 co-authors). 2002. Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. Proc. Natl. Acad. Sci. USA **99**:7548–7553.
Lewin, B. 2000. Genes VII. Oxford University Press, Oxford.
Lewis, E. B. 1978. Gene complex controlling segmentation in Drosophila. Nature **276**:565–570.
Li, Q. M., and S. A. Johnston. 2001. Are all DNA binding and transcription regulation by an activator physiologically relevant? Mol. Cell. Biol. **21**:2467–2474.
Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.
Li, W. W., M. M. Dammerman, J. D. Smith, S. Metzger, J. L. Breslow, and T. Leff. 1995. Common genetic variation in the promoter of the human *apo CIII* gene abolishes regulation by insulin and may contribute to hypertriglyceridemia. J. Clin. Invest. **96**:2601–2605.
Li, X., and M. Noll. 1994. Evolution of distinct developmental functions of three Drosophila genes by acquisition of different *cis*-regulatory regions. Nature **367**:83–87.
Li, Y., J. P. Bernot, C. Illingworth et al. (10 co-authors). 2001. Gene conversion within regulatory sequences generates maize r alleles with altered gene expression. Genetics **159**: 1727–1740.
Liang, Z., and M. D. Biggin. 1998. Eve and ftz regulate a wide array of genes in blastoderm embryos: the selector homeoproteins directly or indirectly regulate most genes in Drosophila. Development **125**:4471–4482.
Lieb, J. D., X. Liu, D. Botstein, and P. O. Brown. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association. Nat. Genet. **28**:327–334.
Liu, T., J. Wu, and F. He. 2000. Evolution of *cis*-acting elements in 5′ flanking regions of vertebrate actin genes. J. Mol. Evol. **50**:22–30.
Locker, J. 2001. Transcription factors. Academic Press, San Diego, Calif.
Long, M., W. Wang, and J. Zhang. 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, *jingwei*, in Drosophila. Gene **238**:135–141.
Loots, G. G., R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science **288**:136–140.
Love, J. J., X. Li, D. A. Case, K. Geise, R. Grosschedl, and P. E. Wright. 1995. Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature **376**:791–795.
Lowe, C. J., and G. A. Wray. 1997. Radical alterations in the roles of homeobox genes during echinoderm evolution. Nature **389**:718–721.
Ludwig, M. Z. 2002. Functional evolution of noncoding DNA. Curr. Opin. Genet. Dev. **12**:634–639.
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.
Ludwig, M. Z., and M. Kreitman. 1995. Evolutionary dynamics

of the enhancer region of *even-skipped* in Drosophila. Mol. Biol. Evol. **12**:1002–1011.
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.
Lufkin, T. 2001. Developmental control by Hox transcriptional regulators and their cofactors. Pp. 215–235 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.
Lutz, B., H. C. Lu, G. Eichele, D. Miller, and T. C. Kaufman. 1996. Rescue of *Drosophila labial* null mutant by the chicken ortholog *Hoxb-1* demonstrates that the function of *Hox* genes is phylogenetically conserved. Genes Dev. **10**:176–184.
Ma, X., D. Yuan, K. Diepold, T. Scarborough, and J. Ma. 1996. The Drosophila morphogenic protein bicoid binds DNA cooperatively. Development **122**:1195–1206.
Maduro, M., and D. Pilgrim. 1996. Conservation of function and expression of *unc-119* from two *Caenorhabditis* species despite divergence of non-coding DNA. Gene **183**:77–85.
Mahmoudi, T., and C. P. Verrijzer. 2001. Chromatin silencing and activation by Polycomb and trithorax group proteins. Oncogene **20**:3055–3066.
Manzanares, M., H. Wada, N. Itasaki, P. A. Trainor, R. Krumlauf, and P. W. Holland. 2000. Conservation and elaboration of *Hox* gene regulation during evolution of the vertebrate head. Nature **408**:854–857.
Margarit, E., A. Guillén, C. Bebordosa, J. Vidal-Taboada, M. Sánchez, F. Ballesta, and R. Oliva. 1998. Identification of conserved potentially regulatory sequences of the *SRY* gene from 10 different species of mammals. Biochem. Biophys. Res. Commun. **245**:370–377.
Markstein, M., and M. Levine. 2002. Decoding *cis*-regulatory DNAs in the Drosophila genome. Curr. Opin. Genet. Dev. **12**:601–606.
Mastick, G. S., R. McKay, T. Oligino, K. Donovan, and A. J. López. 1995. Identification of target genes regulated by homeotic proteins in *Drosophila melanogaster* through genetic selection of Ultrabithorax protein-binding sites in yeast. Genetics **139**:349–363.
Mathias, J. R., H. L. Zhong, H. H. Jin, and A. K. Vershon. 2001. Altering the DNA-binding specificity of the yeast Mat alpha 2 homeodomain protein. J. Biol. Chem. **276**:32696–32703.
Matsuo, Y., and T. Yamazaki. 1984. Genetic analysis of natural populations of *Drosophila melanogaster* in Japan. IV. Natural selection on the inducibility, but not on the structural genes, of amylase loci. Genetics **108**:879–896.
McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351**: 652–654.
McKinney, M. L., and K. J. McNamara. 1991. Heterochrony: the evolution of ontogeny. Plenum Press, New York.
Metherall, J. E., F. P. Gillespie, and B. G. Forget. 1988. Analyses of linked beta-globin genes suggest that nondeletion forms of hereditary persistence of fetal hemoglobin are bona fide switching mutants. Am. J. Hum. Genet. **42**:476–481.
Meyer, C. G., J. May, A. J. Luty, B. Lell, and P. G. Kremsner. 2002. $TNFA^{-308A}$ associated with shorter intervals of *Plasmodium falciparum* reinfections. Tissue Antigens **59**:287–292.
Milo, R., S. Shen-Orr, S. Itzkovitz, D. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: simple building blocks of complex networks. Science **298**:824–827.
Mitsialis, S. A., and F. C. Kafatos. 1985. Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature **317**:453–456.
Miyashita, N. T. 2001. DNA variation in the 5′ upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera*. Mol. Biol. Evol. **18**:164–171.
Mody, M., Y. Cao, Z. Cui et al. (10 co-authors). 2001. Genome-wide gene expression profiles of the developing mouse hippocampus. Proc. Natl. Acad. Sci. USA **98**:8862–8867.
Montano, M. A., C. P. Nixon, T. Ndung'u, H. Bussmann, V. A. Novitsky, D. Dickman, and M. Essex. 2000. Elevated tumor necrosis factor-alpha activation of human immunodeficiency virus type 1 subtype C in southern Africa is associated with an NF-kappaB enhancer gain-of-function. J. Infect. Dis. **181**: 76–81.
Montano, M. A., V. A. Novitsky, J. T. Blackard, N. L. Cho, D. A. Katzenstein, and M. Essex. 1997. Divergent transcriptional regulation among expanding human immunodeficiency virus type 1 subtypes. J. Virol. **71**:8657–8665.
Morishima, A. 1998. Identification of preferred binding sites of a light-inducible DNA-binding factor (MNF1) within 5′-upstream sequence of c4-type *phosphoenolpyruvate carboxylase* gene in maize. Plant Mol. Biol. **38**:633–646.
Mouchel-Vielh, E., M. Blin, C. Rigolot, and J. S. Deutsch. 2002. Expression of a homologue of the *fushi-tarazu* (*ftz*) gene in a cirriped crustacean. Evol. Dev. **4**:76–85.
Müller, G. B., and G. P. Wagner. 1991. Novelty in evolution: restructuring the concept. Annu. Rev. Ecol. Syst. **22**:229–256.
Muller, H. J. 1942. Isolating mechanisms, evolution, and temperature. Biol. Symp. **6**:71–125.
Naganawa, S., H. N. Ginsberg, R. M. Glickman, and G. S. Ginsburg. 1997. Intestinal transcription and synthesis of apolipoprotein AI is regulated by five natural polymorphisms upstream of the apolipoprotein CIII gene. J. Clin. Invest. **99**:1958–1965.
Nakayama, E. E., L. Meyer, A. Iwamoto, A. Persoz, Y. Nagai, C. Rouzioux, J. F. Delfraissy, P. Debre, D. McIlroy, I. Theodorou, T. Shioda, and S. S. Group. 2002. Protective effect of interleukin-4–589T polymorphism on human immunodeficiency virus type 1 disease progression: relationship with virus load. J. Infect. Dis. **185**:1183–1186.
Narlikar, G. J., H.-Y. Fan, and R. E. Kingston. 2002. Cooperation between complexes that regulate chromatin structure and transcription. Cell **108**:475–487.
Naylor, L. H., and E. M. Clark. 1990. d(TG)n.d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. Nucleic Acids Res. **18**:1595–1601.
Neznanov, N., A. Umezawa, and R. G. Oshima. 1997. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. J. Biol. Chem. **272**: 27549–27557.
Nielsen, L. B., D. Kahn, T. Duell, H. U. Weier, S. Taylor, and S. G. Young. 1998. Apolipoprotein B gene expression in a series of human apolipoprotein B transgenic mice generated with recA-assisted restriction endonuclease cleavage-modified bacterial artificial chromosomes. An intestine-specific enhancer element is located between 54 and 62 kilobases 5′ to the structural gene. J. Biol. Chem. **273**:21800–21807.
Nurminsky, D. I., E. N. Moriyama, E. R. Lozovskaya, and D. L. Hartl. 1996. Molecular phylogeny and genome evolution in the *Drosophila virilis* species group: duplications of the *alcohol dehydrogenase* gene. Mol. Biol. Evol. **13**:132–149.
Nurminsky, D. I., M. V. Nurminskaya, D. Deaguiar, and D. L. Hartl. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature **396**:572–575.
Odgers, W. A., M. J. Healy, and J. G. Oakeshott. 1995. Nucleotide polymorphism in the 5′ promoter region of *esterase 6* in *Drosophila melanogaster* and its relationship to enzyme activity variation. Genetics **141**:215–222.
Ohler, U., and H. Niemann. 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet. **17**:56–60.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23**:263–286.

Ohtsuki, S., M. Levine, and H. N. Cai. 1998. Different core promoters possess distinct regulatory activities in the Drosophila embryo. Genes Dev. **12**:547–556.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat. Genet. **32**:261–266.

Onyango, P., W. Miller, J. Lehoczky, C. T. Leung, B. Birren, S. Wheelan, K. Dewar, and A. P. Feinberg. 2000. Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. Genome Res. **10**:1697–1710.

Orphanides, G., T. Lagrange, and D. Reinberg. 1998. The general transcription factors of RNA polymerase II. Genes Dev. **10**:2657–2683.

Orr, H. A., and D. C. Presgraves. 2000. Speciation by postzygotic isolation: forces, genes and molecules. Bioessays **22**:1085–1094.

Osada, S., H. Yamamoto, T. Nishihara, and M. Imagawa. 1996. DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. J. Biol. Chem. **271**:3891–3896.

Paigen, K. 1989. Experimental approaches to the study of regulatory evolution. Am. Nat. **134**:440–458.

Panganiban, G., S. M. Irvine, C. Lowe et al. (14 co-authors). 1997. The origin and evolution of animal appendages. Proc. Natl. Acad. Sci. USA **94**:5162–5166.

Papenbrock, T., R. L. Peterson, R. S. Lee, T. Hsu, A. Kuroiwa, and A. Awgulewitsch. 1998. Murine Hoxc-9 gene contains a structurally and functionally conserved enhancer. Dev. Dyn. **212**:540–547.

Paquette, J., N. Giannoukakis, C. Polychronakos, P. Vafiadis, and C. Deal. 1998. The INS 5′ variable number of tandem repeats is associated with IGF2 expression in humans. J. Biol. Chem. **273**:14158–14164.

Parks, A. L., B. A. Parr, J.-E. Chin, D. S. Leaf, and R. A. Raff. 1988. Molecular analysis of heterochronic changes in the evolution of direct developing sea urchins. J. Evol. Biol. **1**:27–44.

Patrinos, G. P., P. Kollia, A. Loutradi-Anagnostou, D. Loukopoulos, and M. N. Papadakis. 1998. The Cretan type of non-deletional hereditary persistence of fetal hemoglobin [A gamma-158C–>T] results from two independent gene conversion events. Hum. Genet. **102**:629–634.

Petronzelli, F., A. Kimura, P. Ferrante, and M. C. Mazzilli. 1995. Polymorphism in the upstream regulatory region of DQA1 gene in the Italian population. Tissue Antigens **45**:258–263.

Pfister K., K. Paigen, G. Watson, and V. Chapman. 1982. Expression of beta-glucuronidase haplotypes in prototype and congenic mouse strains. Biochem. Genet. **20**:519–536.

Piano, F., M. J. Parisi, R. Karess, and M. P. Kambysellis. 1999. Evidence for redundancy but not trans factor-*cis* element coevolution in the regulation of Drosophila *Yp* genes. Genetics **152**:605–616.

Pinsonneault, J., B. Florence, H. Vaessin, and W. McGinnis. 1997. A model for extradenticle function as a switch that changes Hox proteins from repressors to activators. EMBO J. **16**:2032–2042.

Pirkkala, L., P. Nykanen, and L. Sistonen. 2001. Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. FASEB J. **15**:1118–1131.

Plaza, S., S. Saule, and C. Dozier. 1999. High conservation of *cis*-regulatory elements between quail and human for the *Pax-6* gene. Dev. Genes Evol. **209**:165–173.

Powell, J. R. 1979. Population genetics of Drosophila amylase. II. Geographic patterns in *D. pseudoobscura*. Genetics **92**: 613–622.

Powell, J. R., and J. M. Lichtenfels. 1979. Population genetics of Drosophila amylase. I. genetic control of tissue-specific expression in *D. pseudoobscura*. Genetics **92**:603–612.

Praz, V., R. Perier, C. Bonnard, and P. Bucher. 2002. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expresion data. Nuclic Acids Res. **30**:322–324.

Pugh, B. F. 2001. RNA polymerase II transcription machinery. Pp. 1–16 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.

Purugganan, M. D. 2000. The molecular population genetics of regulatory genes. Mol. Ecol. **9**:1451–1461.

Quiring, R., U. Walldorf, U. Kloter, and W. J. Gehring. 1994. Homology of the *eyeless* gene of Drosophila to the *small eye* gene in mice and *aniridia* in humans. Science **265**:785–789.

Raff, R. A. 1996. The shape of life: genes, development, and the evolution of animal form. The University Of Chicago Press, Chicago.

Raff, R. A., and T. C. Kaufman. 1983. Embryos, genes, and evolution: the developmental-genetic basis of evolutionary change. Macmillan, New York.

Rebeiz, M., N. L. Reeves, and J. W. Posakony. 2002. SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Proc. Natl. Acad. Sci. USA **99**:9888–9893.

Regier, J. C., and N. S. Vlahos. 1988. Heterochrony and the introduction of novel modes of morphogenesis during the evolution of moth choriogenesis. J. Mol. Evol. **28**:19–31.

Reinberg, D., G. Oprhanides, R. Ebright et al. (29 co-authors). 1998. The RNA polymerase II general transcription factors: past, present, and future. Cold Spring Harbor Symp. Quant. Biol. **63**:83–103.

Ren, B., F. Robert, J. J. Wyrick et al. (14 co-authors). 2000. Genome-wide location and function of DNA binding proteins. Science **290**:2306–2309.

Richards, E. J., and S. C. R. Elgin. 2002. Epigenetic codes for heterochromatin silencing: rounding up the usual suspects. Cell **108**:489–500.

Riechmann, J. L., M. Wang, and E. M. Meyerowitz. 1996. DNA-binding properties of Arabidopsis Mads domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. Nucleic Acids Res. **24**:3134–3141.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat Genet. **33**:138–144.

Rinder, H., A. Thomschke, S. Rusch-Gerdes, G. Bretzel, K. Feldmann, M. Rifai, and T. Loscher. 1998. Significance of *ahpC* promoter mutations for the prediction of isoniazid resistance in *Mycobacterium tuberculosis*. Eur. J. Clin. Microbiol. Infect. Dis. **17**:508–511.

Roberts, S. B., N. Segil, and H. Heintz. 1991. Differential phosphorylation of the transcription factor Oct1 during the cell cycle. Science **253**:1022–1026.

Robin, C., R. F. Lyman, A. D. Long, C. H. Langley, and T. F. C. Mackay. 2002. *hairy*: a quantitative trait locus for Drosophila sensory bristle number. Genetics **162**:155–164.

Rockman, M. V., and G. A. Wray. 2002. Abundant raw material for *cis*-regulatory evolution in humans. Mol. Biol. Evol. **19**: 1981–1990.

Romano, L. A., and G. A. Wray. 2003. Conservation of *Endo16* expression in sea urchins despite divergence in both *cis* and *trans*-acting components of transcriptional regulation. Development (in press).

Romey, M. C., C. Guittard, J. P. Chazalette et al. (14 co-authors). 1999. Complex allele [–102T>A+s549r (T>G)] is associated with milder forms of cystic fibrosis than allele s549r[T>G] alone. Hum. Genet. **105**:145–150.

Romey, M. C., N. Pallares-Ruiz, A. Mange, C. Mettling, R. Peytavi, J. Demaille, and M. Claustres. 2000. A naturally occurring sequence variation that creates a YY1 element is associated with increased cystic fibrosis transmembrane conductance regulator gene expression. J. Biol. Chem. **275**: 3561–3567.

Ronshaugen, M., N. McGinnis, and W. McGinnis. 2002. Hox protein mutation and macroevolution of the insect body plan. Nature **415**:914–917.

Ross, J. L., P. P. Fong, and D. R. Cavener. 1994. Correlated evolution of the cis-acting regulatory elements and developmental expression of the Drosophila *Gld* gene in seven species from the subgroup *melanogaster*. Dev. Genet. **15**:38–50.

Rothenburg, S., F. Koch-Nolte, A. Rich, and F. Haag. 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proc. Natl. Acad. Sci. USA **98**:8985–8990.

Ruez, C., F. Payre, and A. Vincent. 1998. Transcriptional control of Drosophila bicoid by Serendipity delta: cooperative binding sites, promoter context, and co-evolution. Mech. Dev. **78**:125–134.

Saccone, G., I. Peluso, D. Artiaco, E. Giordano, D. Bopp, and L. C. Polito. 1998. The *Ceratitis capitata* homologue of the Drosophila sex-determining gene *Sex-lethal* is structurally conserved, but not sex-specifically regulated. Development **125**:1495–1500.

Sackerson, C., M. Fujioka, and T. Goto. 1999. The even-skipped locus is contained in a 16-kb chromatin domain. Dev. Biol. **211**:39–52.

Saito, T., H. M. Lachman, L. Diaz et al. (12 co-authors). 2002. Analysis of monoamine oxidase a (MAOA) promoter polymorphism in Finnish male alcoholics. Psych. Res. **109**: 113–119.

Sandrelli, F., S. Campesan, M. G. Rossetto, C. Benna, E. Zieger, A. Megighian, M. Couchman, C. P. Kyriacou, and R. Costa. 2001. Molecular dissection of the 5′ region of *no-on-transientA* of *Drosophila melanogaster* reveals *cis*-regulation by adjacent dGpi1 sequences. Genetics **157**: 765–775.

Sauer, F., S. K. Hansen, and R. Tjian. 1995. DNA Template requirement and activator-coactivator requirements for transcriptional synergism by *Drosophila bicoid*. Science **270**: 1825–1827.

Scaffidi, P., and M. E. Bianchi. 2001. Spatially precise DNA bending is an essential activity of the Sox2 transcription factor. J. Biol. Chem. **276**:47296–47302.

Scemama, J. L., M. Hunter, J. McCallam, V. Prince, and E. Stellwag. 2002. Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. J. Exp. Zool. **294**:285–299.

Schadt, E. E., S. A. Monks, T. A. Drake et al. (14 co-authors). 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature **422**:297–302.

Schiff, N. M., Y. Feng, J. A. Quine, P. A. Krasney, and D. R. Cavener. 1992. Evolution of the expression of the *Gld* gene in the reproductive tract of Drosophila. Mol. Biol. Evol. **9**:1029–1049.

Schlichting, C. D., and M. Pigliucci. 1998. Phenotypic evolution: a reaction norm perspective. Sinauer Associates, Sunderland, Mass.

Segal, J. A., J. L. Barnett, and D. L. Crawford. 1999. Functional analysis of natural variation in Sp1 binding sites of a TATA-less promoter. J. Mol. Evol. **49**:736–749.

Segil, N., S. B. Roberts, and N. Heitz. 1991. Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. Science **254**:1814–1816.

Seoighe, C., N. Federspiel, T. Jones et al. (29 co-authors). 2000. Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. USA **97**:14433–14437.

Serfling, E., M. Jasin, and W. Schaffner. 1985. Enhancers and eukaryotic gene-transcription. Trends Genet. **1**:224–230.

Shabalina, S. A., and A. S. Kondrashov. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet. Res. **74**:23–30.

Shabalina, S. A., A. Y. Ogurtsov, V. A. Kondrashov, and A. S. Kondrashov. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. **17**:373–376.

Shashikant, C. S., C. B. Kim, M. A. Borbely, W. C. Wang, and F. H. Ruddle. 1998. Comparative studies on mammalian *Hoxc8* early enhancer sequence reveal a baleen whale–specific deletion of a *cis*-acting element. Proc. Natl. Acad. Sci. USA **95**:12364–12369.

Shaw, P. J., N. S. Wratten, A. P. McGregor, and G. A. Dover. 2002. Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. Evol. Dev. **4**:265–277.

Shikama, N., J. Lyon, and N. B. La Thangue. 1997. The p300/CBP family: integrating signals with transcription factors and chromatin. Trends Cell Biol. **7**:230–236.

Shin, H. D., C. Winkler, J. C. Stephens et al. (15 co-authors). 2000. Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. Proc. Natl. Acad. Sci. USA **97**: 14467–14472.

Shore, P., and A. D. Sharrocks. 2001. Regulation of transcription by extracellular signals. Pp. 113–135 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.

Simon, J., M. Peifer, W. Bender, and M. O'Connor. 1990. Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. EMBO J. **9**:3945–3956.

Singh, N., K. W. Barbour, and F. G. Berger. 1998. Evolution of transcriptional regulatory elements within the promoter of a mammalian gene. Mol. Biol. Evol. **15**:312–325.

Singh, N., and F. G. Berger. 1998. Evolution of a mammalian promoter through changes in patterns of transcription factor binding. J. Mol. Evol. **46**:639–648.

Sinha, N. R., and E. A. Kellogg. 1996. Parallelism and diversity in multiple origins of C_4 photosynthesis in the grass family. Am. J. Bot. **83**:1458–1470.

Sinha, S., and M. Tompa. 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res. **30**:5549–5560.

Sjostrand, J. O., A. Kegel, and S. U. Astrom. 2002. Functional diversity of silencers in budding yeasts. Eukaryot. Cell **1**:548–557.

Sjottem, E., C. Andersen, and T. Johansen. 1997. Structural and functional analysis of DNA bending by Sp1 family transcription factors. J. Mol. Biol. **267**:490–504.

Skaer, N., D. Pistillo, and P. Simpson. 2002. Transcriptional heterochrony of *scute* and changes in bristle pattern between two closely related species of blowfly. Dev. Biol. **252**: 31–45.

Skaer, N., and P. Simpson. 2000. Genetic analysis of bristle loss in hybrids between *Drosophila melanogaster* and *D. simulans* provides evidence for divergence of *cis*-regulatory sequences in the *achaete-scute* gene complex. Dev. Biol. **221**:148–167.

Smale, S. T., A. Jain, J. Kaufmann, K. H. Emamai, K. Lo, and I. P. Garraway. 1998. The initiator element: a paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harbor Symp. Quant. Biol. **63**:21–31.

Small, S., A. Blair, and M. Levine. 1992. Regulation of *even-skipped* stripe-2 in the Drosophila embryo. EMBO J. **11**: 4047–4057.

Spek, C. A., R. M. Bertina, and P. H. Reitsma. 1999. Unique distance- and DNA-turn-dependent interactions in the human

protein C gene promoter confer submaximal transcriptional activity. Biochem. J. **340**:513–518.
Stauber, M., H. Jackle, and U. Schmidt-Ott. 1999. The anterior determinant *bicoid* of Drosophila is a derived *Hox* class 3 gene. Proc. Natl. Acad. Sci. USA **96**:3786–3789.
Stauber, M., A. Prell, and U. Schmidt-Ott. 2002. A single *Hox3* gene with composite *bicoid* and *zerknult* expression characteristics in non-Cyclorrhaphan flies. Proc. Natl. Acad. Sci. USA **99**:274–279.
Stern, D. L. 1998. A role of *Ultrabithorax* in morphological difference between Drosophila species. Nature **396**:463–466.
———. 2000. Perspective: evolutionary developmental biology and the problem of variation. Evolution **54**:1079–1091.
Stockhaus, J., U. Schlue, M. Koczor, J. A. Chitty, W. C. Taylor, and P. Westhoff. 1997. The promoter of the gene encoding the C4 form of phoshpoenolpyruvate carboxylase directs mesophyll-specific expression in transgenic C4 *Flaveria* spp. Plant Cell **9**:479–489.
Stone, J. R., and G. A. Wray. 2001. Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol. Biol. Evol. **18**:1764–1770.
Storgaard, T., J. Christensen, B. Aasted, and S. Alexandersen. 1993. Cis-acting sequences in the Aleutian mink disease parvovirus late promoter important for transcription: comparison to the canine parvovirus and minute virus of mice. J. Virol. **67**:1887–1895.
Stormo, G. D. 2000. DNA binding sites: representation and discovery. Bioinformatics **16**:16–23.
Stougaard, J., N. N. Sandal, A. Grøn, A. Kuhle, and K. A. Marcker. 1987. 5′ analysis of the soybean leghaemoglobin *lbc*3 gene: regulatory elements required for promoter activity and organ specificity. EMBO J. **6**:3565–3569.
Streelman, J. T., and T. D. Kocher. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged Tilapia. Physiol. Genomics **9**:1–4.
Struhl, K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell **98**:1–4.
Sucena, E., and D. L. Stern. 2000. Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo*/*shaven-baby*. Proc. Natl. Acad. Sci. USA **97**:4530–4534.
Suske, G. 1999. The Sp-family of transcription factors. Gene **238**:291–300.
Sutton, K. A., and M. Wilkinson. 1997. Rapid evolution of a homeodomain: evidence for positive selection. J. Mol. Evol. **45**:579–588.
Swalla, B. J., and W. R. Jeffery. 1996. Requirement of the *Manx* gene for expression of chordate features in a tailless ascidian larva. Science **274**:1205–1208.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.
Takahashi, A., S. C. Tsaur, J. A. Coyne, and C. I. Wu. 2001. The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **98**:3920–3925.
Takahashi, H., Y. Mitani, G. Satoh, and N. Satoh. 1999. Evolutionary alterations of the minimal promoter for notochord-specific *Brachyury* expression in ascidian embryos. Development **126**:3725–3734.
Takahata, N. 1987. On the overdispersed molecular clock. Genetics **116**:169–179.
Tamarina, N. A., M. Z. Ludwig, and R. C. Richmond. 1997. Divergent and conserved features in the spatial expression of the *Drosophila pseudoobscura esterase-5B* gene and the *esterase-6* gene of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **94**:7735–7741.
Tautz, D. 2000. Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10**:575–579.
Thanos, D., and T. Maniatis. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell **83**:1091–1100.
Theissen, G., A. Becker, A. Di Rosa, A. Kanno, J. T. Kim, T. Munster, K. U. Winter, and H. Saedler. 2000. A short history of MADS-box genes in plants. Plant Mol. Biol. **42**:115–149.
Thompson, J. R., S. W. Chen, L. Ho, A. W. Langston, and L. J. Gudas. 1998. An evolutionary conserved element is essential for somite and adjacent mesenchymal expression of the *Hoxa1* gene. Dev. Dyn. **211**:97–108.
Thurz, M. 2001. Genetic susceptibility in chronic viral hepatitis. Antiviral Res. **52**:113–116.
Ting, C. T., S. C. Tsaur, M. L. Wu, and C. I. Wu. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282**:1501–1504.
Tomarev, S. I., M. K. Duncan, H. J. Roth, A. Cvekl, and J. Piatagorsky. 1994. Convergent evolution of crystallin gene regulation in squid and chicken: the AP-1/ARE connection. J. Mol. Evol. **39**:134–143.
Torchia, J., C. K. Glass, and M. G. Rosenfeld. 1998. Co-activators and co-repressors in the integration of transcriptional responses. Curr. Opin. Cell Biol. **10**:373–383.
Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim. 1995. Disruption of a GATA motif in the *Duffy* gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat. Genet. **10**:224–228.
Trefilov, A., J. Berard, M. Krawczak, and J. Schmidtke. 2000. Natal dispersal in rhesus macaques is related to serotonin transporter gene promoter variation. Behav. Genet. **30**:295–301.
Treisman, J., P. Gönczy, M. Vashishtha, E. Harris, and C. Desplan. 1989. A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell **59**:553–562.
Triezenberg, S. J. 1995. Structure and function of transcriptional activation domains. Curr. Opin. Genet. Dev. **5**:190–196.
Tümpel, S., M. Maconochie, L. M. Wiedemann, and R. Krumlauf. 2002. Conservation and diversity in the cis-regulatory networks that integrate information controlling expression of Hoxa2 in hindbrain and cranial neural crest cells in vertebrates. Dev. Biol. **246**:45–56.
Varga-Weisz, P. 2001. ATP-dependent chromatin remodeling factors: nucleosome shufflers with many missions. Oncogene **20**:3076–3085.
Venter, J. C., M. D. Adams, E. W. Myers et al. (274 co-authors). 2001. The sequence of the human genome. Science **291**:1304–1351.
Vidigal, P., J. J. Gemner, and N. N. Zein. 2002. Polymorphisms in the interleukin-10, tumor necrosis factor-a, and transforming growth factor-b1 genes in chronic hepatitis C patients treated with interferon and ribavarin. J. Hepatol. **36**:271–277.
Vogelauer, M., J. Wu, N. Suka, and M. Grunstein. 2000. Global histone acetylation and deacetylation in yeast. Nature **408**: 495–498.
Von Dassow, G., E. Meir, E. M. Munro, and G. M. Odell. 2000. The segment polarity network is a robust developmental module. Nature **406**:188–192.
Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few duplicate genes. Mol. Biol. Evol. **18**:1283–1292.
Wallace, B. 1963. Genetic diversity, genetic uniformity, and heterosis. Can. J. Genet. Cytol. **5**:239–253.
Walter, J., and M. D. Biggin. 1996. DNA binding specificity of

two homeodomain proteins in vitro and in Drosophila embryos. Proc. Natl. Acad. Sci. USA **93**:2680–2685.
Wang, R. L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. Nature **398**:236–239.
Wang, W., F. G. Brunet, E. Nevo, and M. Long. 2002. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99**:4448–4453.
Wassermann, W. W., M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. **26**:225–228.
Waterston, R. H., K. Lindblad-Toh, E. Birney et al. (219 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.
Wei, Z., L. M. Angerer, M. L. Gagnon, and R. C. Angerer. 1995. Characterization of the *SpHE* promoter that is spatially regulated along the animal-vegetal axis of the sea urchin embryo. Dev. Biol. **171**:195–211.
Weinzierl, R. O. J. 1999. Mechanisms of gene expression. Imperial College Press, London.
West, R. J., R. Yocum, and M. Ptashne. 1984. *Saccharomyces cerevisiae* GAL1-GAL10 divergent promoter region: location and function of the upstream activating sequence UAS_G. Mol. Cell. Biol. **4**:2467–2478.
Wheeler, J. C., K. Shigesada, J. P. Gergen, and Y. Ito. 2000. Mechanisms of transcriptional regulation by Runt domain proteins. Semin. Cell. Dev. Biol. **11**:369–375.
White, K. P., S. A. Rifkin, P. Hurban, and D. S. Hogness. 1999. Microarray analysis of Drosophila development during metamorphosis. Science **286**:2179–2184.
White, R. J. 2001. Gene transcription: mechanisms and control. Blackwell Science, Malden, Mass.
Wilkins, A. S. 1993. Genetic analysis of animal development. Wiley-Liss, Inc., New York.
———. 2002. The evolution of developmental pathways. Sinauer Associates, Sunderland, Mass.
Wilson, A. C. 1975. Evolutionary importance of gene regulation. Stadler Symp. **7**:117–134.
Wilton, A. N., C. C. Laurie-Ahlberg, T. H. Emigh, and J. W. Curtsinger. 1982. Naturally occurring enzyme activity variation in *Drosophila melanogaster*. II. Relationships among enzymes. Genetics **102**:207–221.
Wingender, E., X. Chen, E. Fricke et al. (14 co-authors). 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res. **29**:281–283.
Wolff, C., M. Pepling, P. Gergen, and M. Klingler. 1999. Structure and evolution of a pair-rule interaction element: *runt* regulatory sequences in *D. melanogaster* and *D. virilis*. Mech. Dev. **80**:87–99.
Wolffe, A. P. 1994. Insulating chromatin. Curr. Biol. **4**:85–87.
———. 2001. Chromatin structure and the regulation of transcription. Pp. 35–64 *in* J. Locker, ed. Transcription factors. Academic Press, San Diego, Calif.
Wray, G. A., and A. E. Bely. 1994. The evolution of echinoderm development is driven by several distinct factors. Development **120**(Suppl):97–106.
Wray, G. A., and C. J. Lowe. 2000. Developmental regulatory genes and echinoderm evolution. Syst. Biol. **49**:28–51.
Wray, G. A., and D. R. McClay. 1989. Molecular heterochronies and heterotopies in early echinoid development. Evolution **43**:803–813.
Wright, C. E., F. Haddad, A. X. Quin, P. W. Bodell, and K. M. Baldwin. 1999. In vivo regulation of β-MGC gene in rodent heart: role of T_3 and evidence for an upstream enhancer. Am. J. Physiol. **276**:C883–C891.
Wright, S. 1982. Character change, speciation, and the higher taxa. Evolution **36**:427–443.
Wu, C. Y., and M. D. Brennan. 1993. Similar tissue-specific expression of the *Adh* genes from different Drosophila species is mediated by distinct arrangements of *cis*-acting sequences. Mol. Gen. Genet. **240**:58–64.
Xu, P.-X., X. Zhang, S. Heaney, A. Yoon, A. M. Michelson, and R. L. Maas. 1999. Regulation of *Pax6* expression is conserved between mice and flies. Development **126**:383–395.
Yamamoto, K. R., B. D. Darimont, R. L. Wagner, and J. A. Iñiguez-Lluhí. 1998. Building transcriptional regulatory complexes: signals and surfaces. Cold Spring Harbor Symp. Quant. Biol. **63**:587–598.
Yamamoto, Y., and W. R. Jeffery. 2000. Central role for the lens in cave fish eye degeneration. Science **289**:631–633.
Yan, H., W. S. Yuan, V. E. Velculescu, B. Vogelstein, and K. W. Kinzler. 2002. Allelic variation in human gene expression. Science **297**:1143–1143.
Yu, H., S. H. Yang, and C. J. Goh. 2002. Spatial and temporal expression of the orchid floral homeotic gene DOMADS1 is mediated by its upstream regulatory regions. Plant Mol. Biol. **49**:225–237.
Yuh, C. H., H. Bolouri, and E. H. Davidson. 1998. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. Science **279**:1896–1902.
———. 2001. *Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control. Development **128**:617–629.
Yuh, C.-H., C. T. Brown, C. B. Livi, L. Rowen, P. J. C. Clarke, and E. H. Davidson. 2002. Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin. Dev. Biol. **246**:148–161.
Yun, K. S., and B. Wold. 1996. Skeletal muscle determination and differentiation: story of a core regulatory network and its context. Curr. Opin. Cell Biol. **8**:877–889.
Zeller, R. W., J. D. Griffith, J. G. Moore, C. V. Kirchhamer, R. J. Britten, and E. H. Davidson. 1995. A multimerizing transcription factor of sea urchin embryos capable of looping DNA. Proc. Natl. Acad. Sci. USA **92**:2989–2993.
Zerucha, T., T. Stuhmer, G. Hatch, B. K. Park, Q. M. Long, G. Y. Yu, A. Gambarotta, J. R. Schultz, J. L. R. Rubenstein, and M. Ekker. 2000. A highly conserved enhancer in the *Dlx5*/*Dlx6* region is the site of cross-regulatory interactions between *Dlx* genes in the embryonic forebrain. J. Neuro. **20**:709–721.
Zhu, J., and M. Q. Zhang. 1999. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. Bioinformatics **15**:607–611.
Zuckerkandl, E. 1963. Perspectives in molecular anthropology. Pp. 256–274 *in* A. Rich and N. Davidson, eds. Structural chemistry and molecular biology. W.H. Freeman, San Francisco.

William Jeffery, Associate Editor

Accepted April 1, 2003