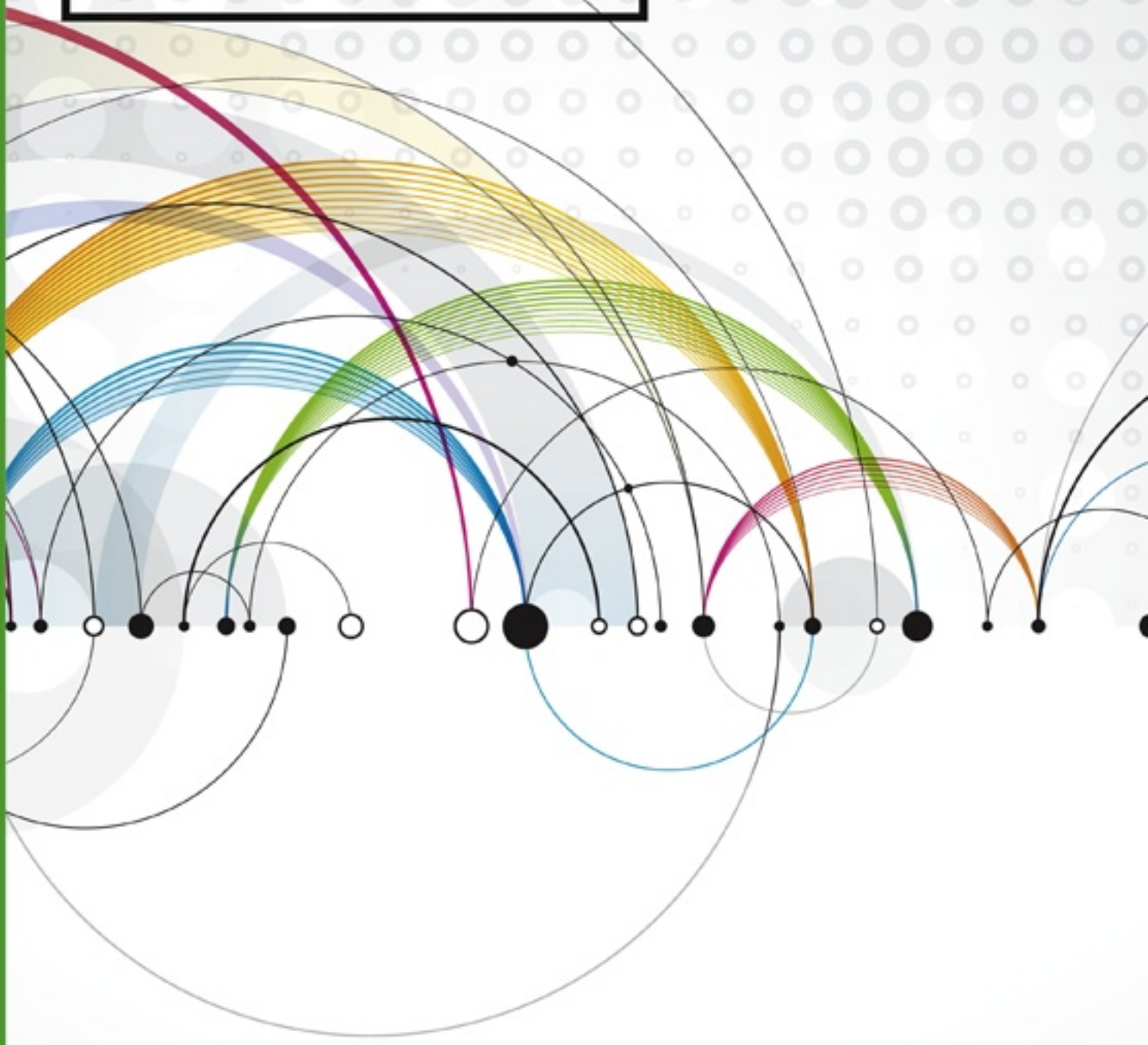
Trends in Genetics January 2013 Vol. 29 No. 1, pp. 1–58 ISSN 0168-9525

Trends in Genetics



Cracking the *cis*-regulatory code

Cell PRESS

Review

Cell PRESS

Deciphering the transcriptional *cis*-regulatory code

J. Omar Yáñez-Cuna*, Evgeny Z. Kvon*, and Alexander Stark

Research Institute of Molecular Pathology (IMP), 1030 Vienna, Austria

Information about developmental gene expression resides in defined regulatory elements, called enhancers, in the non-coding part of the genome. Although cells reliably utilize enhancers to orchestrate gene expression, a *cis*-regulatory code that would allow their interpretation has remained one of the greatest challenges of modern biology. In this review, we summarize studies from the past three decades that describe progress towards revealing the properties of enhancers and discuss how recent approaches are providing unprecedented insights into regulatory elements in animal genomes. Over the next years, we believe that the functional characterization of regulatory sequences in entire genomes, combined with recent computational methods, will provide a comprehensive view of genomic regulatory elements and their building blocks and will enable researchers to begin to understand the sequence basis of the *cis*-regulatory code.

Regulatory information is encoded in defined DNA sequence elements

During development, a single precursor cell (the fertilized egg) gives rise to a complex multicellular organism comprising a large variety of cell types. This process is heritable and repeats generation after generation according to a developmental program that is encoded in the genome. This program governs the dynamic regulation of gene expression in response to environmental and developmental stimuli and thus determines the differentiation of cell types, their morphologies and functions, and other biological processes during development.

The information about when and where a gene is going to be expressed resides in defined genomic elements called, largely synonymously, enhancers, *cis*-regulatory elements, or *cis*-regulatory modules (CRMs; Figure 1a,b). Studies on important developmental regulators, such as *decapentaplegic* (*dpp*) or *even-skipped* (*eve*) in *Drosophila*, showed that their expression is controlled by multiple enhancers whose spatiotemporal activities contribute additively to gene expression (reviewed in [1]; Figure 1a), suggesting that the regulation of gene expression is organized in a modular fashion with enhancers being the modular units (thus the term 'CRM'; see Glossary). The contribution of individual enhancers to the overall expression pattern of the gene, and the necessity for the enhancers, can be demonstrated by BAC reporter constructs, as done recently for *Blimp-1*, for example [2]. However, in addition to cases of additive enhancer function, many genes have several seemingly redundantly acting enhancers, possibly to ensure robustness of gene expression (reviewed in [3]).

Enhancers contain multiple binding sites for sequence-specific transcription factors (TFs), and the combined regulatory input from all bound TFs (and the cofactors they recruit) results in the spatiotemporal-specific activation of target gene transcription (Figure 1b). When placed out of their endogenous genomic context, enhancers recapitulate endogenous TF binding [4,5], DNA and histone modifications [4,6], and cell type-specific enhancer activities, that is, endogenous gene expression patterns (reviewed in [7]; Figure 1a). This suggests that the *cis*-regulatory information required for the creation of these patterns is encoded within the DNA sequences of the enhancers (reviewed in [8]). Indeed, changes to the primary DNA sequence have been associated with differences in gene expression and TF binding between human individuals [9,10] and between closely related species [4,11–13]. Because such alterations can be heritable, they might contribute to phenotypic variation within a species (e.g., [10]) and morphological changes during evolution [14].

Cells are able to interpret enhancer sequences reliably, but understanding enhancer sequences and predicting their function has remained one of the greatest challenges in biology. It is attractive to speculate that regulatory information is encoded by a defined set of rules regarding enhancer sequence composition and organization. In analogy to the 'genetic code' that allows the translation of protein-coding DNA sequences into amino acid sequences, we refer to rules that would allow the functional interpretation of enhancer sequences as the '*cis*-regulatory code', irrespective of the concrete nature of such rules.

Here, we discuss and summarize approaches and findings relevant to unraveling the *cis*-regulatory code, concentrating on studies in fruit flies. We specifically focus on the DNA sequence characteristics of enhancers, whereas mechanistic aspects regarding the cellular machinery used to read and interpret *cis*-regulatory information is beyond the scope of this review.

Understanding the *cis*-regulatory code

The comparison between protein-coding and regulatory DNA sequences, that is, between the genetic and the regulatory code, is interesting and instructive. Proteins are encoded in open reading frames (ORFs) that are

Corresponding author: Stark, A. (stark@starklab.org)

Keywords: enhancers; *cis*-regulatory elements; regulatory code; transcription factors; regulatory genomics; gene regulation.

* These authors contributed equally.

Glossary

ChIP-on-chip/Chip-seq: high-throughput techniques to identify DNA binding sites of a given protein genome-wide. These techniques have been applied to proteins that bind DNA directly or indirectly, including TFs, cofactors, and histones (reviewed in [44]).

***Cis*-regulatory information**: we use the term '*cis*-regulatory information' in an abstract sense as the information required to explain or (re-) create the activity pattern of an enhancer in a given cellular environment. Because enhancers typically retain their activity patterns independently of their sequence contexts (e.g., in reporter constructs), their DNA sequences contain all *cis*-regulatory information.

Core promoter: region around a transcription start site (TSS; typically +/–40 bp) that contains motifs for DNA-binding proteins involved in the recruitment, assembly, initiation, pausing and/or stalling, and elongation of transcription by RNA polymerase II (reviewed in [107]). Core promoters are typically not active without enhancers and are often also termed 'minimal promoters'.

Cross-validation: an approach to obtain an unbiased estimate of prediction performance on unseen objects. This is achieved by dividing the curated data into two non-overlapping and independent sets, which are exclusively used for training or testing, respectively (training set versus test set). The importance of cross-validation and of carefully separating training and test sets is crucial, because results might otherwise be overestimated (see discussion in [67]). This is especially true for complex models with a high number of parameters, or when the number of objects is limited, as is often the case in biology. Further care is warranted if objects are related non-trivially, for example by remote sequence similarity or by temporal and spatial trends of gene expression and/or enhancer activity.

Developmental kernel: core subcircuit of a gene regulatory network. Developmental kernels that regulate the development of essential conserved body structure have been found to be shared across evolutionarily distant species, in which homologous TFs form equivalent regulatory core connections (reviewed in [87]).

(Transcriptional) Enhancers: regulatory regions in the genome that control the transcription of their target genes, typically in a cell type-specific manner, yet largely independently of their genomic sequence context and their position relative to their target genes (see main text for exceptions). The term 'enhancer' was originally coined based on viral SV40 DNA sequences that enhanced the transcription of a beta-globin reporter gene from a core (or minimal) promoter in ectopic assays [94]. CRMs or *cis*-regulatory regions are often used largely synonymously, although these terms more strongly emphasize the modular architecture of gene regulation with several enhancers per gene, each contributing different aspects of the expression pattern of the genes (e.g., the seven stripes of *eve*; Figure 1a) and also include more general regulatory activities. Even though 'enhancer' originally described the functional property of viral or animal DNA sequences independent of mechanistic aspects, the term is also partly used less strictly for genomic regions presumed to have regulatory roles due to their association with TFs or certain histone modifications irrespective of whether the ability of the region to activate and/or enhance transcription has been established.

Machine learning: computational methods designed for the classification of objects by identifying consistent patterns or correlations within the data. These approaches are typically divided into two groups according to whether they use curated data for training (learning model parameters; supervised learning) or not (unsupervised learning). Due to the growing availability of large amounts of information to analyze, machine-learning approaches are becoming increasingly useful in molecular biology, such as in the case of enhancer and TF binding site prediction (see [108]).

Support vector machine (SVM): a widely used supervised machine-learning approach that separates objects into groups by a hyperplane in high-dimensional feature space (see [108]).

TF motif: typically, a short and degenerate DNA sequence pattern preferentially or specifically recognized by a TF, generally represented as IUPAC 'consensus motif' or PWM [33].

Transcription factors (TFs): proteins that bind DNA in a sequence-specific manner and activate or repress transcription of a target gene, usually via the recruitment of cofactors. The term 'transcription factor' also includes general transcription factors (GTFs), which typically do not bind DNA directly and are part of the basal transcriptional machinery that, together with RNA polymerase II, forms the pre-initiation complex. Animals typically contain hundreds of TFs, which are increasingly being catalogued by online databases, such as FlyTF [106].

defined by start and stop codons in between which the ungapped sequential occurrence of 61 nucleotide triplets or 'codons' determines the linear amino acid sequence of the protein. The genetic code therefore amounts to a simple mapping of all 4^3 (= 64) possible triplets redundantly to the different amino acids and stop codons.

In contrast to the well-understood and simple nature of the genetic code, it has remained unclear whether a regulatory code exists in terms of a set of rules that predict enhancer function from the sequence and explain the sequence characteristics of enhancers active in a specific cell type or across cell types (Figure 1c). Enhancer sequences contain short DNA words or 'motifs' that are recognized and bound by TFs, which contribute activating and repressing cues via the recruitment of cofactors with diverse functions. However, in contrast to codons in ORFs, the motifs are of variable lengths and are interspersed by gaps of seemingly neutral sequence. Their order and arrangement appears to be variable and differs between enhancers of similar functions (e.g., between muscle enhancers; Figure 2a; [15,16]). It also appears to be flexible within individual enhancers, because motif rearrangements during evolution (e.g., [17] and references therein) or in experimental tests [18] often do not impair enhancer function (Figure 2b). Also in contrast to ORFs, the boundaries of enhancers are less well defined (often through rather coarse trimming regulatory regions to the shortest sequences that are still functional), such that, for example, functional binding sites can be found outside these 'minimal enhancers' (see [19]).

Although such complexities might argue against the existence of a regulatory code, recent work indicates that enhancers with similar functions share sequence characteristics and do not function by entirely different means (e.g., [5,20–22]. This suggests that rules exist that are applicable across functionally similar enhancers and might allow the identification of important sequence features, such as the occurrence of TF motifs, potentially in specific combinations or relative arrangements (Figure 1c). Below, we discuss efforts and progress towards this goal and explain why we believe that an understanding of *cis*-regulatory codes for defined cellular or developmental contexts is in reach.

Enumerating the parts

The classical approach to study the rules that govern enhancer function has been the exhaustive characterization of individual enhancers, such as *eve* stripe 2 (Figure 1a,b; see [23] and references therein) and *sparkling* in *Drosophila* [18,24] or the interferon-beta enhancer in mammals [25]). These studies established much of what is known today about the genomic properties of enhancers and suggested that the main 'building blocks' of the *cis*-regulatory code are TF motifs; the presence of certain TF motif combinations and partly their specific arrangements (sometimes referred to as 'grammar') appear to govern enhancer function. In addition, other sequence features, such as nucleotide and di-nucleotide composition, influence properties of the DNA and nucleosome occupancy, which in turn influence TF binding [26,27]. The identification of all building blocks and the potential rules regarding their arrangement has become the biggest endeavor towards understanding the *cis*-regulatory code.

Recently, the characterization of individual enhancers has been scaled up substantially by combining DNA synthesis with barcode labeling to allow exhaustive mutational tests of defined enhancers in yeast and mammalian cells

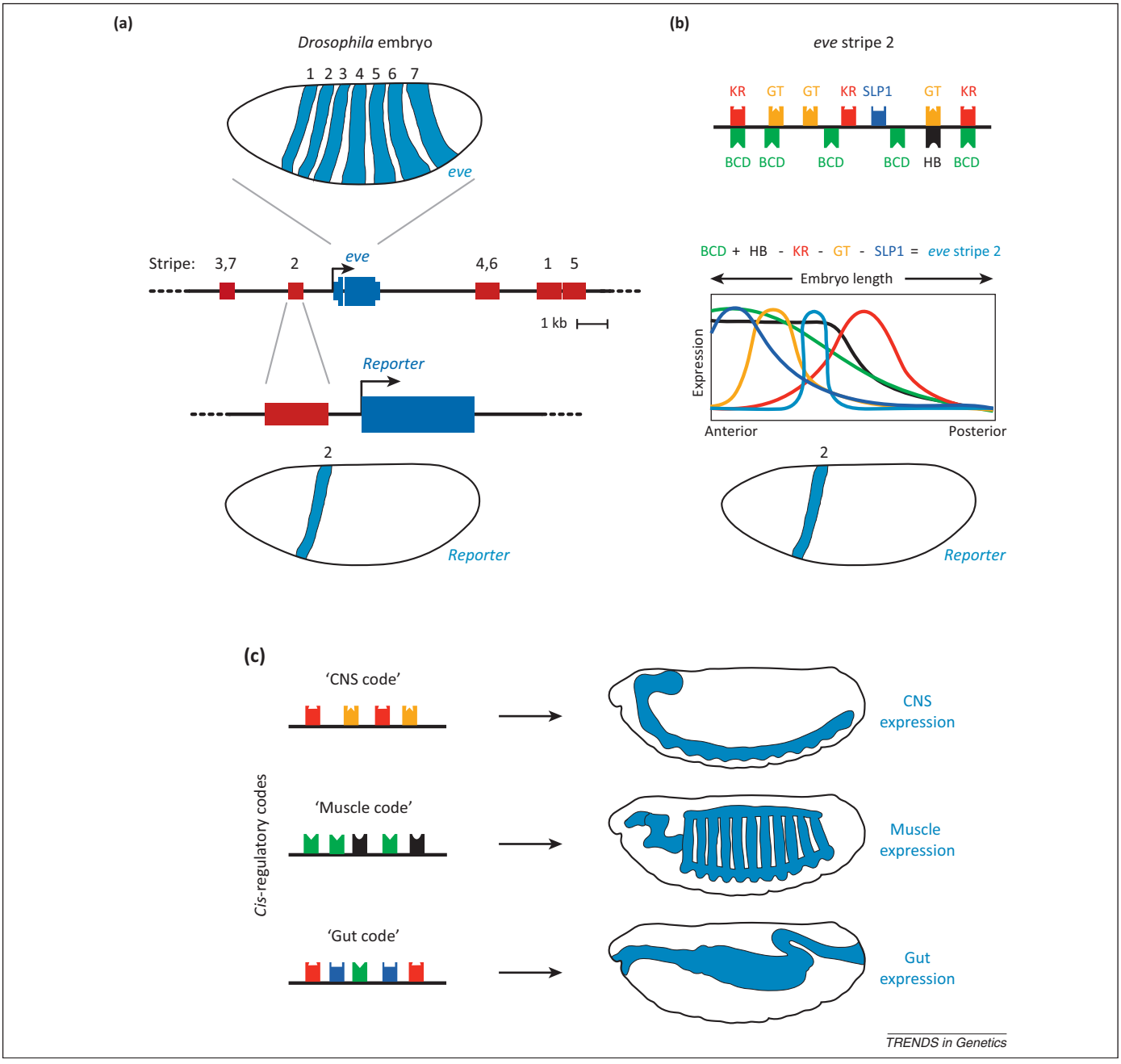


Figure 1. Transcriptional enhancers. **(a)** General properties of transcriptional enhancers [or *cis*-regulatory modules (CRMs)] illustrated using the regulatory locus of the pair-rule gene *even-skipped* (*eve*). *Eve* is expressed in seven distinct anterior–posterior stripes in the early *Drosophila* embryo, which result from the additive input of five different enhancers (red blocks) that lie at variable positions with respect to the transcription start site of *eve*. Each enhancer is active in one or two stripes and recapitulates this activity even when placed in transgenic reporter constructs (shown is the *eve* stripe 2 enhancer). **(b)** Enhancer sequences contain transcription factor (TF) motifs that act as TF binding sites. Shown are known TF motifs (colored blocks; top) within the *eve* stripe 2 enhancer for the activators Bicoid (BCD; green blocks) and Hunchback (HB; black) and the repressors Giant (GT; yellow), Kruppel (KR; red), and Sloppy paired 1 (SLP1; blue). The TF protein concentration along the anterior–posterior axis determines TF binding to the enhancer and the combined input of all bound TFs results in the activity of the enhancer: the activators provide broad activating input, whereas the three repressors shape the borders of stripe 2 (see [23] and references therein). **(c)** Functionally related enhancers active in certain cell types or tissues might share important features regarding motif composition and organization, that is, a cell type-specific *cis*-regulatory code. Shown are cartoons illustrating such putative codes for the *Drosophila* embryonic central nervous system (CNS), muscle, or gut. (Embryo figures were adapted from [119].)

([28–30]; Figure 3a). More generally, the study of genomic regulatory elements has greatly benefited from methods that make use of microarray and next-generation sequencing technologies, allowing genome-wide assays with vastly increased statistical power to study the building blocks of the regulatory code.

Building block I: TF motifs

TF motifs capture the DNA sequence preferences of TFs and are therefore typically short and degenerate, which can be represented by IUPAC consensus sequences or more flexibly by position-specific weight matrices (PWMs). TF motifs are found in enhancer sequences and their importance for enhancer function has often been demonstrated by the combination of genetic and biochemical approaches. For example, in-depth analysis identified motifs for five TFs in *eve* stripe 2, arguably the most well-characterized developmental enhancer (Figure 1a,b; see [23] and references therein). Each occurrence of a motif in an enhancer sequence is typically scored according to how well it

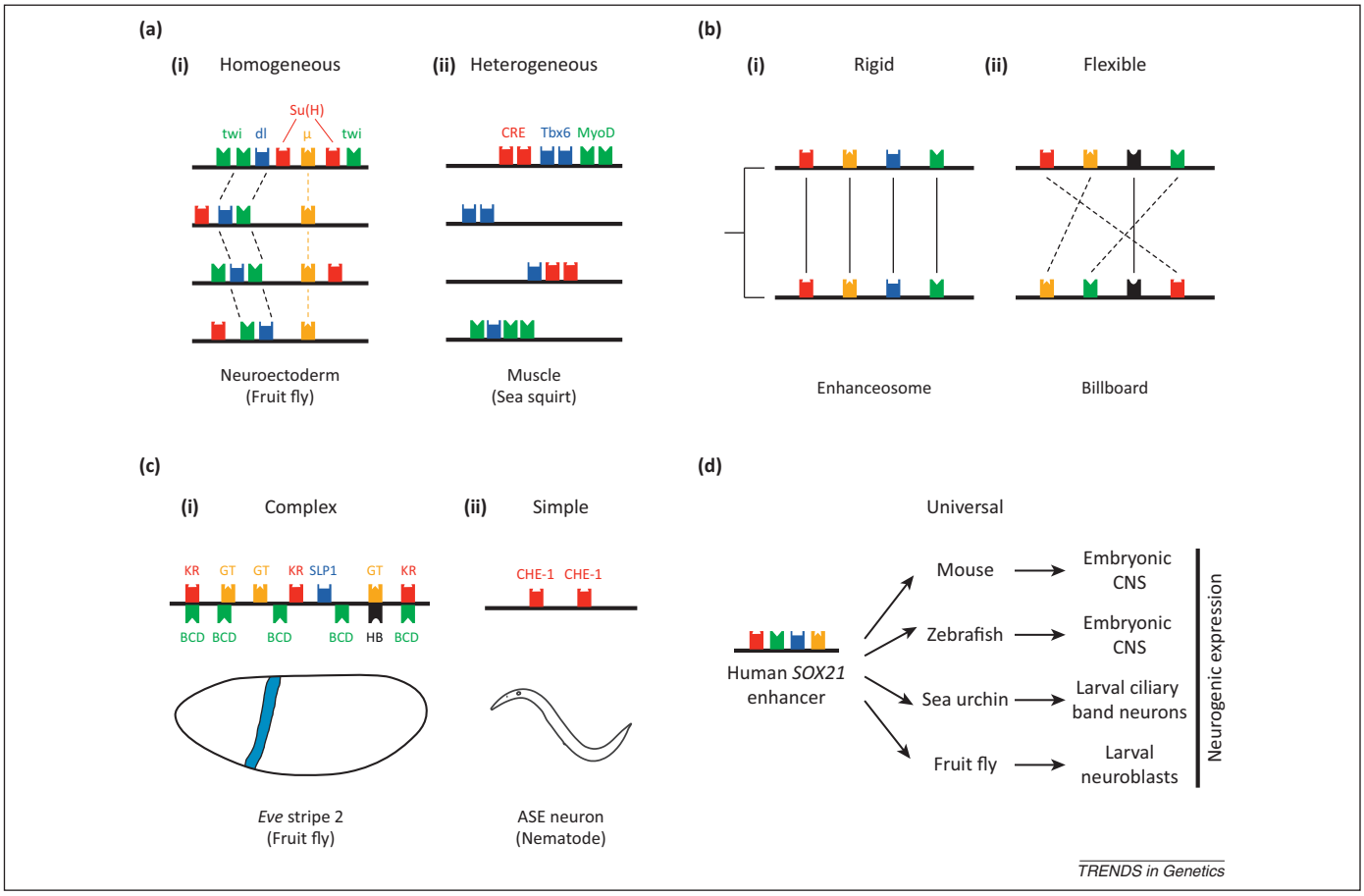


Figure 2. Enhancer structure and properties of the *cis*-regulatory code. **(a)** Homogeneous versus heterogeneous. Shown is a schematic representation of enhancers active in the neurogenic ectoderm of early *Drosophila* embryos (i) [21] and in muscle cells of *Ciona* embryos (ii) [15]. The overall motif composition and structure of neurogenic ectoderm enhancers (NEE) is similar ('homogeneous code'), and it shows several organizational constraints, such as the presence of Suppressor of Hairless (Su(H)) motifs or even the precise relative positioning of Twist (twi), Dorsal (dl), and mystery (μ) motifs [21]. By contrast, *Ciona* muscle enhancers appear highly dissimilar without any apparent constraints regarding the spacing, orientation, or even presence of cAMP response element (CRE), T-box 6 (Tbx6), and Myogenic Differentiation (MyoD) motifs ('heterogeneous code'; [15]). **(b)** Rigid versus flexible. Shown are two extreme examples of functional constraints on enhancer architecture: In the 'enhanceosome' type (i), enhancers are characterized by precise spacing, positioning, and orientation of motifs, which are functionally required and maintained in evolution ('rigid code'). In 'billboard'-type enhancers (ii), motif presence rather than precise motif positioning appears to be important, and the less constrained architecture often means that sequence conservation is more difficult to detect ('flexible code'; see main text and [7] for details). **(c)** Complex versus simple. Enhancers involved in developmental patterning often appear more complex compared with enhancers that regulate downstream gene batteries in defined cell types or that regulate downstream gene batteries. For example, the *eve* stripe 2 enhancer contains multiple binding sites for at least five TFs (i) (see Figure 1 for details), whereas enhancers active in ASE (Amphid neurons, single ciliated endings) gustatory neuron of *Caenorhabditis elegans* receive prominent input only from the TF abnormal CHEmotaxis (CHE-1) (ii) [20]. Note that the known motifs shown here are not sufficient for activity in either case. **(d)** Universality. The enhancer of the human *SOX21* gene is able to drive neurogenic expression in embryos across phyla, including different vertebrates, the tunicate *Ciona*, and even *Drosophila* larvae [85]. Abbreviation: CNS, central nervous system.

matches the degeneracy pattern of the PWM, distinguishing, for example, high- and low-affinity motifs, which both can be important and serve different functions [31,32]. Given the importance of TF motifs, many approaches have been developed for their discovery, often based on the enrichment of motifs in enhancer regions [33,34].

TF motif discovery using gene coexpression

It has been intriguing to speculate that genes that are functionally related or coexpressed share *cis*-regulatory motifs recognized by a common set of regulators (Figures 1c and 2a). This approach has been successful in finding novel motifs even when only a handful of functionally characterized sequences were available; motifs involved in the terminal differentiation of neurons in *Caenorhabditis elegans* (e.g., [20]; Figures 2c and 3b) or in the activity of neurogenic ectoderm enhancers in *Drosophila melanogaster* [22] were identified this way. With new methods such as DNA microarrays and RNA-seq, it has become possible to determine the expression level of all genes in the genome. Such an approach was used to reveal overrepresented motifs in regions upstream of sets of coexpressed genes genome-wide (reviewed in [34,35]), a strategy that could be extended to larger regulatory regions around genes [36,37].

This approach also revealed the so-called 'TAGteam motif' in the upstream regions of early zygotic genes in *Drosophila* embryos [38], leading to the identification of Zelda as the corresponding TF [39]. The TAGteam motif is also enriched for TF binding sites, based on ChIP experiments in early embryos [40], and its presence correlates with TF binding across species [11]. Zelda binds to essentially all early embryonic enhancers [2,41,42] and appears to be important for enhancer function and binding of other TFs, and might also be a general activator of the early zygotic genome ([5,39,41]; Figure 3d). Interestingly, this includes the well-characterized *eve* stripe 2 enhancer [43], where the important role of Zelda remained unnoticed despite almost 30 years of intensive study.

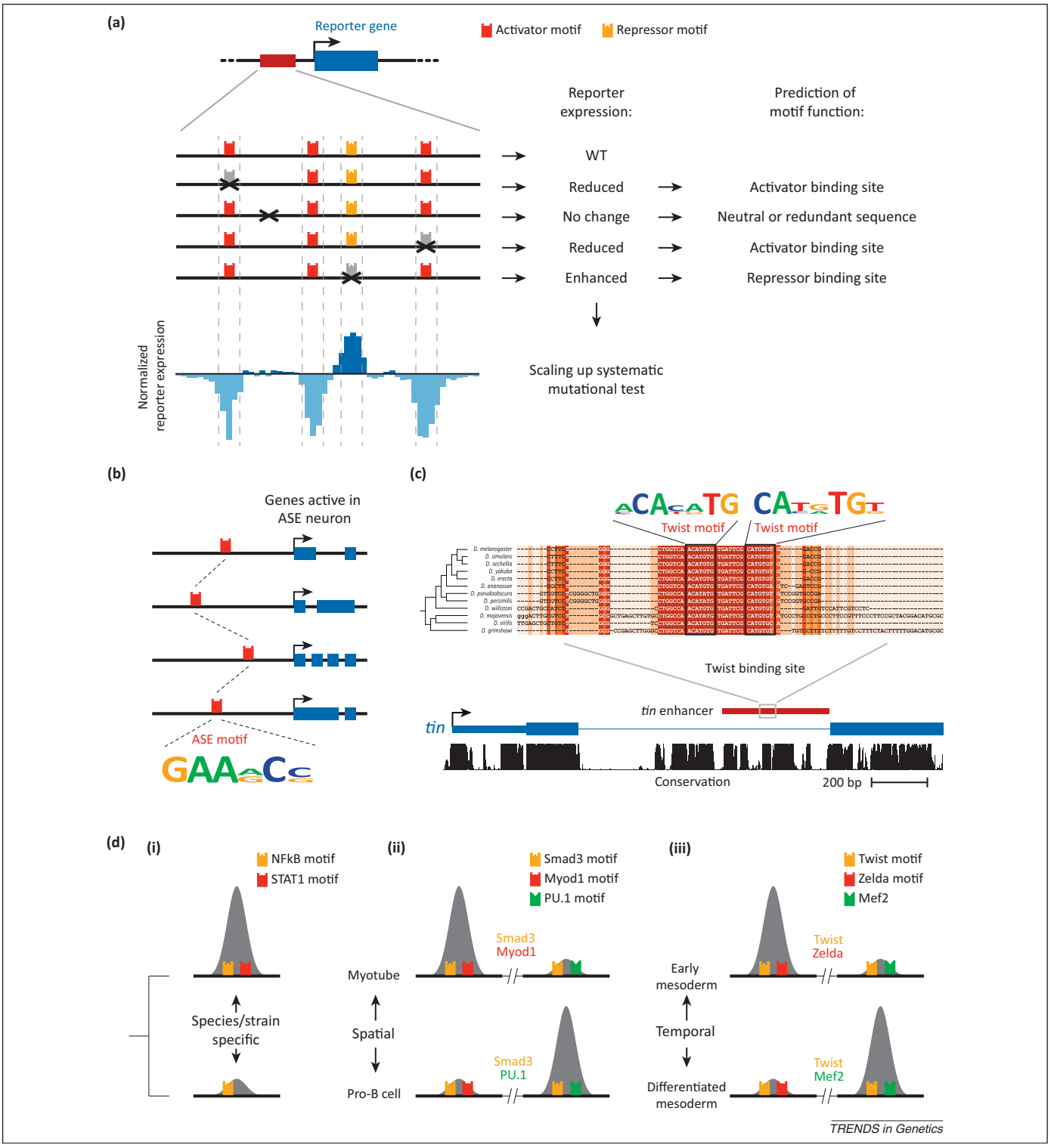


Figure 3. Dissecting enhancers: towards the building blocks of the *cis*-regulatory code. **(a)** The fine dissection of individual enhancers allows the identification of important binding sites by assessing the changes in expression after mutations that disrupt selected motifs (top) or systematically cover the enhancer sequence (bottom). Changes in reporter expression indicate the disruption of sequence elements that function to activate or repress transcription. **(b)** Regions upstream of coexpressed genes in the same species can be enriched in certain sequence motifs, which allows their discovery, e.g., the CHE-1 motif upstream of genes coexpressed in ASE neurons in *Caenorhabditis elegans* [20]. **(c)** Sequence conservation between species can be used to identify regulatory regions, individual transcription factor (TF) binding sites, and regulatory motif types. Shown are two highly conserved Twist binding sites within a mesodermal enhancer in the *tinman* (*tin*) locus. **(d)** Potential partner TFs can be overrepresented and discovered in context-specific TF binding sites, as identified by ChIP-on-Chip and ChIP-seq. For example, variation in nuclear factor κB (NFκB) occupancy between human individuals correlates with the presence of signal transducer and activator of transcription 1 (STAT1) motifs (i) [9]; the tumor growth factor (TGF)-beta signaling effector Mothers against decapentaplegic homolog 3 (Smad3) binds near motifs for the pioneer TFs Myogenic Differentiation 1 (MyoD1) and Purine-rich box-1 (PU.1) in myotubes and near Smad3 motifs in Pro-B cells, respectively (ii) [56], and stage-specific binding of Twist during *Drosophila* embryo development has been predicted using motifs for stage-specific partner TFs, such as Zelda or myocyte enhancer factor 2 (Mef2) (iii) [5].

TF motif discovery using in vitro *and* in vivo *TF binding*

ChIP-on-chip and ChIP-seq (reviewed in [44]) are now frequently used to map TF and cofactor binding as well as histone or DNA modifications genome-wide and have enabled the characterization of regulatory regions in entire genomes (Box 1). Importantly, ChIP pinpoints the *in vivo* binding sites of regulators, which occur at variable locations with respect to their target genes and are otherwise difficult to find. Such binding sites usually allow the discovery of the DNA binding motif of the respective regulator [34], and we expect the power of these approaches to increase with the resolution of DNase-seq (e.g., [45]) and the recently developed ChIP-exo method [46], which can both delineate the actual binding site footprints with nucleotide resolution genome-wide.

TF motifs from ChIP studies usually agree well with motifs obtained by *in vitro* binding assays (such as EMSA, SELEX, PBMs, or bacterial-one-hybrid) for many TFs in yeast [47], flies [48], and mammals [49–51]. In addition, yeast-one-hybrid allows the identification of the TF that might bind to a selected putative regulatory DNA sequence, which is now being enabled by genome-wide libraries of *C. elegans*, *Drosophila*, and *Arabidopsis* TFs [52–54]. In the near future, these efforts will reveal a comprehensive picture of many more TF motifs.

TF motif discovery using conservation

As functional genomic elements are typically under negative evolutionary selection and are shared in different species, sequence conservation has been used as a means to discover functional elements (Figure 3c). Sequence comparisons across related species allowed the *de novo* discovery of motif types by their high average conservation across all motif occurrences genome-wide or in selected genomic regions (e.g., upstream of genes) in yeast, flies, and vertebrates (see [55]). Although individual motif occurrences are frequently diverged, the average motif conservation of many occurrences can provide a robust readout for functionality that also accounts for potential artifacts during genome sequence alignment and boosts the signal-to-noise ratio, which is especially relevant if only a few or closely related species are to be compared (see [7] for a discussion).

Building block II: TF cooperativity

Interestingly, *in vivo* TF binding studies using ChIP revealed two surprises: a TF binds to only a small fraction of all motif occurrences in the genome, and the binding sites differ between contexts (i.e., tissue or cell type), suggesting that the TF motif alone is not sufficient to direct *in vivo* binding (e.g., [5,56–58] and references therein).

Some of these differences may be due to the presence of additional factors required for TF binding *in vivo*. Indeed, the comparison of binding sites for a single TF across different contexts has revealed motifs for partner factors that are predictive of binding ([5]; Figure 3d). This requirement is also supported by findings that TF binding across different species and between individuals within one species correlates with the presence of partner TF motifs near the binding site sequences [9,11,12,59] (Figure 3d). Partner factors and the corresponding motif combinatorics are

Box 1. Approaches to identify enhancers

In contrast to yeast and worms, where regulatory regions are usually located just upstream of the genes, enhancers in higher eukaryotes can be found at variable distances from their target genes, both upstream and downstream or even within the genes themselves (Figure 1a, main text), which makes their identification difficult. Several experimental and computational approaches have been developed to identify or predict enhancers in the genome (reviewed in [55]).

Transgenic reporter assays

Reporter assays are currently the only method to test and characterize enhancers functionally. Enhancers predicted by the other methods discussed here are typically tested using reporter assays. Recent efforts are increasing the throughput of reporter assays to enable the screening of more candidate sequences [2,28–30,109] (E. Kvon and A. Stark, unpublished). Functionally characterized enhancers are often collected in databases, such as REDfly [110] for flies, or VISTA enhancer browser for mice and humans [63].

Chromatin profiling

Genomics approaches enable enhancer prediction from the genome-wide profiling of enhancer-associated histone marks [e.g., histone H3 lysine 4 mono-methylation (H3K4me1) or histone H3 lysine 27 acetylation (H3K27ac)] cofactor [e.g., CREB binding protein (CBP)/P300] binding sites, or chromatin accessibility data (reviewed in [44,111]).

TF binding sites

Similar to chromatin profiling, ChIP-on-chip and ChIP-seq profiling of TF binding sites can be used to identify potential enhancers genome-wide (reviewed in [44,111]). Because most TF binding sites identified using ChIP seem to have no apparent effect on gene expression [40], TF clustering can be used to improve the predictive power of ChIP methods (e.g., [2,16,40]).

Computational enhancer prediction

The computational prediction of enhancers from genome sequences typically leverages one of two statistical signals (or a combination of both): the local enrichment of TF motifs within short regions (TF motif clustering) and the evolutionary conservation of TF motifs (phylogenetic footprinting and related approaches).

TF motif clustering

Based on the observation that known enhancers contain several sequence motifs for one or more TFs, one approach to predict new enhancers searches for clusters of TF motifs in genome sequences, partly requiring that motifs are conserved or occur in certain combinations or arrangements (reviewed in [55]). Interestingly, however, even regions with many TF motifs are not always functional (e.g., [112]).

Phylogenetic footprinting

Regulatory regions and individual TF motif instances can be predicted by their significant conservation in otherwise less conserved sequences. These islands of conservation can be interpreted as evolutionary or phylogenetic 'footprints' that arise from the negative selection of functionally important sequences (Figure 3c, main text; see [55]). Especially due to the recent and ongoing sequencing of closely related fly [113], nematode [114], or vertebrate species [115], phylogenetic footprinting has allowed the prediction of regulatory regions and even individual TF motif instances in yeast, flies, and vertebrates, and is expected to improve with the increasing number of sequenced genomes (see [115–117] for an analysis of the scaling of statistical power of comparative genome sequence analysis). This success is remarkable, because enhancers can maintain function despite largely diverged primary sequences due to divergent or compensatory changes, which means that individual enhancers might not always be detected by conservation-based approaches (e.g., [118]; reviewed in [7]).

attractive because they have the potential to explain cell type-specific binding and function of more broadly expressed TFs and can be considered as a first step towards understanding the sequence basis of gene regulation.

Building block III: additional enhancer sequence features

The detailed analysis of individual enhancers, such as the enhancer *sparkling*, which drives the expression of *shaven* [the *Drosophila* homolog of vertebrate paired box 2 (*Pax2*)] in cone cells of the developing *Drosophila* eye, has revealed the complexity of enhancer sequences [18,24]: even after the detailed dissection of all binding sites for relevant TFs, the sequences between these were found to be essential [18]. This suggests that much of what is needed for enhancer activity is still unknown, even for well-characterized enhancers. For *sparkling*, many of these essential sequences contain motifs that likely correspond to previously unknown TF binding sites, whereas others might be independent of TFs: the 'remote control element' or RCE [18] for example appears to mediate enhancer–promoter interactions. In addition, genome-wide nucleosome maps (e.g., [26]) have identified several sequences that correlate with nucleosome occupancy, and nucleosome disfavoring sequences can influence TF binding and enhancer activity [60]. Similarly, mammalian regulatory elements can contain CpG islands that can be methylated according to the cellular context, thus adding an extra level of regulation [6].

Assembling the building blocks: 'Enhancer Grammar'

Detailed dissections of individual enhancers have also provided insights into the relative arrangements of TF motifs: within the *eve* stripe 2 enhancer, for example, the precise positioning and orientation of TF binding sites appears to be less important than the combined input of the TFs. Such flexibility has also been observed for other developmental enhancers (e.g., [18,24,61]) and might be a general property of the regulatory code (Figure 2b; see also below). In sharp contrast to this model of largely unconstrained enhancer architecture (billboard model), the enhancer of the interferon-beta gene in mammals requires the precise composition, spacing, and order of binding sites for the proper assembly of protein factors into a functional regulatory complex, exemplifying a highly constrained enhancer (enhanceosome model; see [7] for a recent comparison of both models). Although probably rare, other examples of rigid enhanceosomes might exist, for example in 'ultraconserved regions' [62], many of which function as developmental enhancers [63].

Predicting regulatory function from the building blocks

If gene expression is determined by a regulatory code that can be generalized across different enhancers with similar functions, it should be possible to learn rules (e.g., about the presence of TF motifs or binding sites) from known enhancers (training set) and predict the functionality of previously unseen sequences (test set) (Figure 4a). If truly independent sequences are used during training and testing, such 'cross-validated' predictions allow powerful conclusions to be drawn; in addition to inferring the functionality of novel sequences, they implicitly test the importance of the combined features and allow the weighing of each feature: successful predictions imply that the features capture regulatory information that holds true across different enhancers.

This strategy has been applied to functionally characterized sets of regulatory sequences in several systems. In yeast, for example, in which regulatory regions are generally located just upstream of genes, it has been possible to predict expression categories for genes based on the motif content of the upstream regions of the genes [64–67]. In higher eukaryotes, where the location of enhancers is variable and largely still unknown, the earliest approaches to predict developmental enhancers in *Drosophila* were based on TF motif clustering, an established property of enhancers [68,69]. More recently, *in vivo* binding sites of TFs and cofactors have been used as landmarks to identify regulatory sequences (e.g., [16,70]; Box 1). This allowed both the classification of enhancers into groups of specific activities based on the temporal profiles of TF binding [16] and the prediction of both cofactor and TF tissue- and/or cell type-specific binding from the motif content of the binding site sequences [5,71]. Using the sequences of known tissue-specific enhancers as training sets, the prediction of novel enhancers with tissue-specific activities has also been achieved [36,37,72–75], suggesting that a regulatory code exists.

Towards a mechanistic and quantitative understanding of enhancers

Several modeling approaches based on thermodynamics or logistic regression have been used to predict expression and to seek a quantitative and mechanistic understanding of enhancer function and gene expression from first principles (reviewed in [76]; Figure 4b). Such approaches are informed by biological and biophysical knowledge, and attempt to model the binding of activators and repressors (TFs) to the DNA, the recruitment of intermediate proteins, such as cofactors, or mediator components, polymerase recruitment, and transcription [77,78]. Most models are based on the assumption that TFs bind to enhancers at equilibrium, such that TF occupancy can be computed as a function of binding affinities and TF protein concentration. Through weights that model the activating or repressing function of each TF, polymerase occupancy and gene expression are computed (reviewed in [76]). In addition to predicting novel enhancers [79], modeling gene expression in strictly cross-validated settings can assess understanding of enhancers via the performance of the respective model (see [67,80] for a discussion on cross-validation and parameter robustness during modeling of transcriptional regulation). In addition, through measuring the contributions of different model parameters, light can be shed on the importance of the biological factors these parameters represent: homo- or heterotypic cooperative interactions between TFs, transcriptional synergy, and short-range repression have been modeled to assess their relative importance for enhancer activity ([31,78,79,81,82]; reviewed in [76]). For the anterior–posterior axis specification during *Drosophila* embryonic development, for example, modeling suggested that the repressors Kruppel

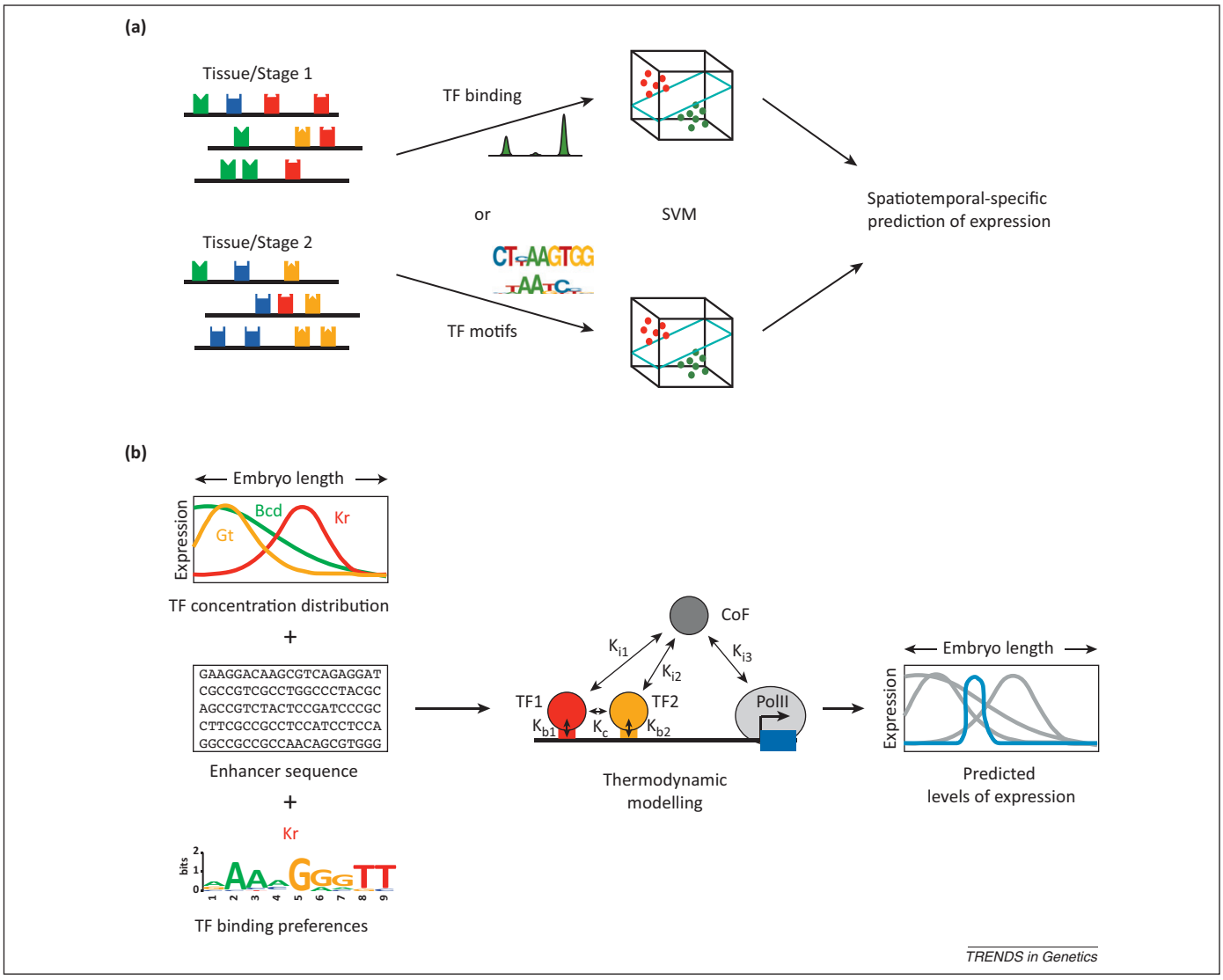


Figure 4. Predicting enhancer activity. **(a)** Predicting enhancer activity in a classification framework. Features that discern two classes of functionally similar enhancers (e.g., active in 'Tissue/Stage 1' versus 'Tissue/Stage 2') can be learned from a training set to classify unseen enhancers in a test set using standard machine-learning tools in a cross-validated setting [e.g., support-vector machines (SVMs)]. Input features might be the TF binding profile of the enhancers (e.g., [16]) or their sequence motif content (e.g., [5,71–73], the weighted importance of which can be interpreted as 'regulatory rules'. **(b)** Schematic representation of the (thermodynamic) modeling of enhancer activity and gene expression based on TF concentrations and binding affinities (based on [120]). Modeling typically assumes conditions of thermodynamic equilibrium and requires as input the TF concentrations, restricting it largely to well-characterized biological systems, such as the *Drosophila* early blastoderm embryo. As in (a), parameters are fit on known enhancers to predict the activity of unseen enhancers (e.g., using cross-validation). Abbreviations: CoF, Co-factor; K_{xx}, Equilibrium constants; Kr, Kruppel; Bcd, Bicoid; Gt, Giant, PolII, RNA polymerase II. Based on [120].

(KR) and Hunchback (HB) function via short-range repression and that transcriptional synergy (i.e., the simultaneous interaction of several TFs with the transcriptional machinery), as well as cooperative binding, are important [82]. While the discriminatory classification approaches described in the last section treat the regulatory *trans*-environment of each cell implicitly and require training for each system (i.e., cell or activity pattern), thermodynamic models, once trained, are potentially applicable to other cell types for which reliable information about TF concentration and activity is available.

Universality of the *cis*-regulatory code

In contrast to the universality of the genetic code, the evolutionary distance between species that are able to correctly interpret the enhancer sequences of each other is more limited and difficult to assess. This is because cell types and their complements of *trans*-acting regulators (which interpret the *cis*-regulatory sequences) are often restricted to certain phyla and are themselves subject to evolutionary change (see [83]).

Homologous *cis*-regulatory sequences have often been found to function correctly even when diverged beyond recognizable sequence similarity between species as divergent as *Drosophila*, sepsids or *Anopheles*, or between human and fish (reviewed in [7]). This is partly due to the flexibility and redundancy of regulatory sequences and to compensatory changes [84]. Some sequences have been found to even function across phyla, such as the enhancer of the human SRY (sex determining region Y)-box 21 gene (*SOX21*; a soxB2 class gene), which is highly conserved from human to *Nematostella* at the sequence level and active in the central nervous system of zebrafish embryos and in *Drosophila* larval neuroblasts, indicating that a 'neuronal regulatory state' is deeply conserved [85] (Figure 2d). Another example of extremely conserved

regulatory connections are the so-called 'developmental kernels' (regulatory core connections), which include eye development [86], muscle and heart development [87], and growth regulation (Myc; [88]).

In contrast to the deep conservation of enhancer function, individual regulatory connections or inputs from certain TFs diverge frequently. For example, genomic binding sites of the CCAAT/enhancer-binding protein alpha (CEBPA) and hepatocyte nuclear factor 4-alpha (HNF4A) [13] in vertebrate liver have diverged significantly between different species.

The *cis*-regulatory code and evolution

The modularity of enhancer function and the flexibility and redundancy of the *cis*-regulatory code, especially in comparison with the genetic code, can explain both its functional robustness and the apparent ease with which sequence changes can alter gene expression. This has important consequences for evolutionary dynamics, and changes in the transcriptional regulation of genes are considered to be one of the major drivers for morphological evolution [14]. In particular, the contribution of several independent cell type-specific enhancers to the overall expression pattern of a gene means that gain- or loss-of-function mutations can have cell type-specific rather than pleiotropic (and likely detrimental) effects [14]. Indeed, examples such as pelvic reduction in sticklebacks [89] or the loss of sensory vibrissae and penile spines in humans [90] have been directly associated with the loss of tissue-specific enhancer function. In addition, the combinatorial and partly redundant regulatory input of several TFs into one enhancer means that changes can also occur at the level of individual enhancers: *de novo* creation of TF motifs can add additional TF inputs, thereby expanding the domains of activity. For example, the gain of new binding sites for conserved regulators of wing development in the enhancer of the pigmentation gene *yellow* created a new wing spot in *Drosophila biarmipes* [91]. If this affects the expression of TFs themselves, the entire *trans*-regulatory environment of cells can be altered, affecting many downstream target genes and causing cell-fate changes (e.g., [92]). Such scenarios have been suggested as part of a general model of *cis*-regulatory changes driving evolution [14], in which alterations at the DNA sequence level led to the creation or loss of TF binding, thereby changing the wiring of regulatory connections within existing *trans*-regulatory environments.

Many of the insights gained from studies of regulation across different species will be relevant for the understanding of disease-causing regulatory mutations [93], which abound and lead to etiological changes in gene expression that are only now beginning to be understood.

The road ahead: enhancer elements, their genomic context, and enhancer–promoter interactions

Transcriptional enhancers have fascinated and puzzled researchers since their initial discovery more than 30 years ago [94]. Although many characteristics of enhancers have been uncovered since, there is still no comprehensive picture of the necessary sequence features for even the most well-studied enhancers, and the *de novo* creation of an enhancer from non-functional sequence by the addition of such features has yet to be achieved. Similarly incomplete is the picture of the abundance and location of enhancers in animal genomes: the spatiotemporal activity pattern is known only for a small number of enhancers and the vast majority of the non-coding genome remains functionally uncharacterized in any animal.

However, as outlined here, individual research groups and international consortia are in the process of determining the binding preferences and the *in vivo* binding sites for increasing numbers of TFs and cofactors, as well as the genome-wide distribution of histone modifications and chromatin states [95–97]. In addition, spatiotemporal gene expression and enhancer activity patterns are systematically being determined in *Drosophila* embryos [2,98] (E. Kvon and A. Stark, unpublished). In the upcoming years, we foresee that these efforts will provide an unprecedented level of functional annotation of regulatory elements, which will enable computational approaches similar to those described above to reveal sequence elements that are important and required for enhancer function. We speculate that it might be possible to determine combinations of elements that are sufficient for the function of certain classes of enhancers, allowing the *de novo* creation of functional enhancers from non-functional sequences.

In addition to the relation between defined enhancer sequences and functions, exciting progress is also being made in answering two related long-standing questions: the functional role of the genomic context of an enhancer and how enhancer–promoter interactions are specified and established across large genomic distances. Ever since the discovery of position-effect variegation in *Drosophila* [99], it has been apparent that proximity to certain genomic contexts, such as heterochromatin or Polycomb recruitment sites, can influence enhancer activity and gene expression independently of the enhancer sequence (reviewed in [100,101]). We expect ongoing work to determine enhancer activity at different genomic positions (e.g., by transposon mediated enhancer traps in mice [102]) together with the concurrent determination of chromatin properties at such sites (e.g., chromatin modifications, insulator, or cofactor binding sites [95–97,103]) to provide significant insight into this question. Similarly, recently developed methodology to assess systematically spatial proximity of different genomic regions (see overview in [104,105]) enables new approaches to study how enhancer–promoter contacts are specified and established across large genomic distances, a question that has remained mysterious.

Methods to study regulation systematically across entire genomes will make the upcoming years undoubtedly an exciting time to study the genomic regulatory elements and the principles of the transcriptional regulatory code.

Acknowledgments

We would like to thank Hannes Tkadletz (IMP/IMBA Graphics Department) for help with the figures and the anonymous reviewers for their helpful comments. We apologize to the many scientists the work of whom we could not cite owing to formal restrictions. Our work is supported by a European Research Council (ERC) Starting Grant from the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 242922 awarded to A.S. and by the

Austrian Ministry for Science and Research through the GEN-AU Bioinformatics Integration Network III. Basic research at the Research Institute of Molecular Pathology (IMP) is supported by Boehringer Ingelheim GmbH.

References

1 Zeitlinger, J. and Stark, A. (2010) Developmental gene regulation in the era of genomics. *Dev. Biol.* 339, 230–239
2 Kvon, E.Z. *et al.* (2012) HOT regions function as patterned developmental enhancers and have a distinct *cis*-regulatory signature. *Genes Dev.* 26, 908–913
3 Lagha, M. *et al.* (2012) Mechanisms of transcriptional precision in animal development. *Trends Genet.* http://dx.doi.org/10.1016/j.tig.2012.03.006
4 Wilson, M.D. *et al.* (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* 322, 434–438
5 Yanez-Cuna, J.O. *et al.* (2012) Uncovering *cis*-regulatory sequence requirements for context specific transcription factor binding. *Genome Res.* http://dx.doi.org/10.1101/gr.132811.111
6 Lienert, F. *et al.* (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* 43, 1091–1097
7 Meireles-Filho, A.C.A. and Stark, A. (2009) Comparative genomics of gene regulation-conservation and divergence of *cis*-regulatory information. *Curr. Opin. Genet. Dev.* 19, 565–570
8 Istrail, S. and Davidson, E.H. (2005) Logic functions of the genomic *cis*-regulatory code. *Proc. Natl. Acad. Sci. U.S.A.* 102, 4954–4959
9 Kasowski, M. *et al.* (2010) Variation in transcription factor binding among humans. *Science* 328, 232–235
10 Reddy, T.E. *et al.* (2012) Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* 22, 860–869
11 Bradley, R.K. *et al.* (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* 8, e1000343
12 He, Q. *et al.* (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat. Genet.* 43, 414–420
13 Schmidt, D. *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036–1040
14 Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25–36
15 Brown, C.D. *et al.* (2007) Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* 317, 1557–1560
16 Zinzen, R.P. *et al.* (2009) Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* 462, 65–70
17 Ludwig, M.Z. *et al.* (2005) Functional evolution of a *cis*-regulatory module. *PLoS Biol.* 3, e93
18 Swanson, C.I. *et al.* (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev. Cell* 18, 359–370
19 Ludwig, M.Z. *et al.* (2011) Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genet.* 7, e1002364
20 Etchberger, J.F. *et al.* (2007) The molecular signature and *cis*-regulatory architecture of a *C. elegans* gustatory neuron. *Genes Dev.* 21, 1653–1674
21 Erives, A. and Levine, M. (2004) Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3851–3856
22 Markstein, M. *et al.* (2004) A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* 131, 2387–2394
23 Andrioli, L.P.M. *et al.* (2002) Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms. *Development* 129, 4931–4940
24 Swanson, C.I. *et al.* (2011) Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr. Biol.* 21, 1186–1196
25 Thanos, D. and Maniatis, T. (1995) Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* 83, 1091–1100
26 Mavrich, T.N. *et al.* (2008) Nucleosome organization in the *Drosophila* genome. *Nature* 453, 358–362
27 Parker, S.C.J. *et al.* (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324, 389–392
28 Melnikov, A. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277
29 Patwardhan, R.P. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* 30, 265–270
30 Sharon, E. *et al.* (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* http://dx.doi.org/10.1038/nbt.2205
31 Zinzen, R.P. *et al.* (2006) Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr. Biol.* 16, 1358–1365
32 Parker, D.S. *et al.* (2011) The *cis*-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Sci. Signal.* 4, ra38
33 Stormo, G.D. and Zhao, Y. (2010) Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.* 11, 751–760
34 Stormo, G.D. (2010) Motif discovery using expectation maximization and Gibbs' sampling. *Methods Mol. Biol.* 674, 85–95
35 MacIsaac, K.D. and Fraenkel, E. (2006) Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput. Biol.* 2, e36
36 Aerts, S. *et al.* (2010) Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in *Drosophila* uncovers a regulatory basis for sensory specification. *PLoS Biol.* 8, e1000435
37 Warner, J.B. *et al.* (2008) Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods* 5, 347–353
38 ten Bosch, J.R. *et al.* (2006) The TAGteam DNA motif controls the timing of *Drosophila* pre-blastoderm transcription. *Development* 133, 1967–1977
39 Liang, H-L. *et al.* (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in *Drosophila*. *Nature* 456, 400–403
40 Li, X-Y. *et al.* (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* 6, e27
41 Nien, C-Y. *et al.* (2011) Temporal coordination of gene networks by Zelda in the early *Drosophila* embryo. *PLoS Genet.* 7, e1002339
42 Harrison, M.M. *et al.* (2011) Zelda binding in the early *Drosophila* melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet.* 7, e1002266
43 Struffi, P. *et al.* (2011) Combinatorial activation and concentration-dependent repression of the *Drosophila* even skipped stripe 3+7 enhancer. *Development* http://dx.doi.org/10.1242/dev.065987
44 Maston, G.A. *et al.* (2012) Characterization of enhancer function from genome-wide analyses. *Annu. Rev. Genomics Hum. Genet.* http://dx.doi.org/10.1146/annurev-genom-090711-163723
45 Neph, S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90
46 Rhee, H.S. and Pugh, B.F. (2011) Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* 147, 1408–1419
47 Badis, G. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* 32, 878–887
48 Noyes, M.B. *et al.* (2008) A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.* 36, 2547–2560
49 Badis, G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723
50 Jolma, A. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 20, 861–873
51 Berger, M.F. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266–1276
52 Gaudinier, A. *et al.* (2011) Enhanced Y1H assays for *Arabidopsis*. *Nat. Methods* 8, 1053–1055

53 Deplancke, B. *et al.* (2006) A gene-centered *C. elegans* protein–DNA interaction network. *Cell* 125, 1193–1205
54 Hens, K. *et al.* (2011) Automated protein–DNA interaction screening of *Drosophila* regulatory elements. *Nat. Methods* 8, 1065–1070
55 Hardison, R.C. and Taylor, J. (2012) Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat. Rev. Genet.* 13, 469–483
56 Mullen, A.C. *et al.* (2011) Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* 147, 565–576
57 Trompouki, E. *et al.* (2011) Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* 147, 577–589
58 Heinz, S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589
59 Zheng, W. *et al.* (2010) Genetic analysis of variation in transcription factor binding in yeast. *Nature* 464, 1187–1191
60 Khoueiry, P. *et al.* (2010) A *cis*-regulatory signature in ascidians and flies, independent of transcription factor binding sites. *Curr. Biol.* 20, 792–802
61 Ho, M.C.W. *et al.* (2009) Functional evolution of *cis*-regulatory modules at a homeotic gene in *Drosophila*. *PLoS Genet.* 5, e1000709
62 Bejerano, G. *et al.* (2004) Ultraconserved elements in the human genome. *Science* 304, 1321–1325
63 Visel, A. *et al.* (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* 40, 158–160
64 Bussemaker, H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–171
65 Pilpel, Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159
66 Beer, M.A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell* 117, 185–198
67 Yuan, Y. *et al.* (2007) Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* 3, e243
68 Berman, B.P. *et al.* (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. U.S.A.* 99, 757–762
69 Schroeder, M.D. *et al.* (2004) Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* 2, E271
70 Visel, A. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854–858
71 Lee, D. *et al.* (2011) Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* 21, 2167–2180
72 Narlikar, L. *et al.* (2010) Genome-wide discovery of human heart enhancers. *Genome Res.* 20, 381–392
73 Busser, B.W. *et al.* (2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet.* 8, e1002531
74 Rouault, H. *et al.* (2010) Genome-wide identification of *cis*-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proc. Natl. Acad. Sci. U.S.A.* http://dx.doi.org/10.1073/pnas.1002876107
75 Van Loo, P. *et al.* (2008) ModuleMiner – improved computational detection of *cis*-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol.* 9, R66
76 Segal, E. and Widom, J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.* 10, 443–456
77 Reinitz, J. *et al.* (2003) Transcriptional control in *Drosophila*. *Complexus* 1, 54–64
78 Janssens, H. *et al.* (2006) Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene. *Nat. Genet.* 38, 1159–1165
79 Kazemian, M. *et al.* (2010) Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biol.* 8, e1000456
80 Dresch, J.M. *et al.* (2010) Thermodynamic modeling of transcription: sensitivity analysis differentiates biological mechanism from mathematical model-induced effects. *BMC Syst. Biol.* 4, 142
81 Fakhouri, W.D. *et al.* (2010) Deciphering a transcriptional regulatory code: modeling short-range repression in the *Drosophila* embryo. *Mol. Syst. Biol.* 6, 341
82 He, X. *et al.* (2010) Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput. Biol.* 6, e1000935
83 Arendt, D. (2008) The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* 9, 868–882
84 Ludwig, M.Z. *et al.* (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567
85 Royo, J.L. *et al.* (2011) Transphyletic conservation of developmental regulatory state in animal evolution. *Proc. Natl. Acad. Sci. U.S.A.* http://dx.doi.org/10.1073/pnas.1109037108
86 Halder, G. *et al.* (1995) Induction of ectopic eyes by targeted expression of the eyeless gene in *Drosophila*. *Science* 267, 1788–1792
87 Davidson, E.H. and Erwin, D.H. (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311, 796–800
88 Brown, S.J. *et al.* (2008) Evolution of the holozoan ribosome biogenesis regulon. *BMC Genomics* 9, 442
89 Chan, Y.F. *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327, 302–305
90 McLean, C.Y. *et al.* (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471, 216–219
91 Gompel, N. *et al.* (2005) Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433, 481–487
92 Frankel, N. *et al.* (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* 474, 598–603
93 Visel, A. *et al.* (2009) Genomic views of distant-acting enhancers. *Nature* 461, 199–205
94 Banerji, J. *et al.* (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308
95 The modENCODE Consortium *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* http://dx.doi.org/10.1126/science.1198374
96 Gerstein, M.B. *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE Project. *Science* http://dx.doi.org/10.1126/science.1196914
97 ENCODE Project Consortium *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
98 Tomancak, P. *et al.* (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 3, research0088.1-0088.14
99 Muller, H. (1930) Types of visible variations induced by X-rays in *Drosophila*. *J. Genet.* 22, 299–334
100 Spitz, F. and Furlong, E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626
101 Lelli, K.M. *et al.* (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* http://dx.doi.org/10.1146/annurev-genet-110711-155437
102 Ruf, S. *et al.* (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat. Genet.* 43, 379–386
103 Filion, G.J. *et al.* (2010) Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* http://dx.doi.org/10.1016/j.cell.2010.09.009
104 Sanyal, A. *et al.* (2011) Chromatin globules: a common motif of higher order chromosome structure? *Curr. Opin. Cell Biol.* 23, 325–331
105 de Wit, E. and de Laat, W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26, 11–24
106 Adryan, B. and Teichmann, S.A. (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* 22, 1532–1533
107 Juven-Gershon, T. *et al.* (2008) The RNA polymerase II core promoter – the gateway to transcription. *Curr. Opin. Cell Biol.* 20, 253–259
108 Tarca, A.L. *et al.* (2007) Machine learning and its applications to biology. *PLoS Comput. Biol.* 3, e116
109 Pennacchio, L.A. *et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502
110 Gallo, S.M. *et al.* (2010) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res.* http://dx.doi.org/10.1093/nar/gkq999

111 Buecker, C. and Wysocka, J. (2012) Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet.* 28, 276–284
112 Berman, B.P. *et al.* (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* 5, R61
113 Drosophila 12 Genomes Consortium *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218
114 Kumar, S. *et al.* (2012) 959 Nematode Genomes: a semantic wiki for coordinating sequencing projects. *Nucleic Acids Res.* 40, D1295–D1300
115 Lindblad-Toh, K. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482
116 Eddy, S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* 3, e10
117 Stark, A. *et al.* (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450, 219–232
118 Blow, M.J. *et al.* (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* 42, 806–810
119 Campos-Ortega, J.A. and Hartenstein, V. (1997) *The Embryonic Development of Drosophila Melanogaster*, Springer
120 Segal, E. *et al.* (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 451, 535–540