MODES OF TRANSCRIPTIONAL REGULATION

In pursuit of design principles of regulatory sequences

Michal Levo and Eran Segal

Abstract | Instructions for when, where and to what level each gene should be expressed are encoded within regulatory sequences. The importance of motifs recognized by DNA-binding regulators has long been known, but their extensive characterization afforded by recent technologies only partly accounts for how regulatory instructions are encoded in the genome. Here, we review recent advances in our understanding of regulatory sequences that influence transcription and go beyond the description of motifs. We discuss how understanding different aspects of the sequence-encoded regulation can help to unravel the genotype–phenotype relationship, which would lead to a more accurate and mechanistic interpretation of personal genome sequences.

Genome-wide association studies
(GWASs). Genome-wide studies that are designed to identify genetic associations with an observable trait, disease or condition.

Expression quantitative trait locus
(eQTL). A locus at which genetic variation is associated with variation in gene expression levels.

Since the discovery of regulatory DNA sequences in the 1960s[1], a main research focus has been to unravel how regulatory instructions are encoded within these sequences. Advances in our ability to decipher regulatory sequences hold promise to substantially improve our understanding of fundamental biological processes such as differentiation and development, as changes to these sequences are associated with species-specific morphology and were shown to alter properties of expression patterns that are required for development[2,3]. Importantly, deviations from the desired expression pattern (which stem from misregulation) often underlie the formation of diseases, as observed in recent genome-wide association studies (GWASs)[4–7] and studies of cancer[8,9], in which many gene expression changes that are characteristic of the disease state have been linked to changes in regulatory regions.

The basic paradigm for the execution of regulatory instructions encoded within the DNA accounts for the timely binding of regulatory proteins — mainly, transcription factors (TFs) — to promoters or enhancers, which in turn prompt the recruitment of the transcription machinery to the core promoter, resulting in transcription initiation and the formation of robust expression patterns[10,11]. In recent years, a surge of new high-throughput and quantitative technologies have greatly advanced our understanding of transcription and our ability to characterize the regulatory sequences involved. These technologies enable us to efficiently produce and manipulate a large number of DNA sequences, to carry out measurements of DNA binding by various regulators (BOX 1), and to measure the expression driven by thousands of native and synthetically designed regulatory sequences (BOXES 2,3).

Here, we review our current understanding of regulatory sequences (for example, promoters and enhancers), focusing primarily on recent studies. We first discuss the current characterization of the main 'building blocks' of regulatory sequences — namely, TF binding sites (TFBSs) — and the degree to which knowledge of their properties allows the prediction of TF binding in cells. We further explore the various combinations of TFBSs within regulatory sequences and how gene expression depends on the properties of these combinations (for example, the composition and arrangement of TFBSs). We then go beyond the regulatory elements and discuss the possible effects of the sequence context of these elements that are mediated, for example, by the chromatin landscape and by DNA structure. Finally, we discuss the incorporation of accumulating knowledge on regulatory sequences into GWASs and expression quantitative trait locus (eQTL) analyses, as well as the potential of such approaches for identifying causal single-nucleotide polymorphisms (SNPs) and for predicting the expression driven by personal genome sequences.

Regulatory building blocks

TFBSs are considered the core building blocks of regulatory sequences. These sequences are fairly short (6–12 bp) and have distinct specificity for DNA-binding molecules — namely, TFs — that play a part in regulating the expression of the associated genes when they are bound. The characterization of TFBSs has considerably improved in recent years through the development of several high-throughput methods.

Department of Molecular Cell Biology, and Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.
Correspondence to E.S.
e-mail:
eran.segal@weizmann.ac.il

doi:10.1038/nrg3684
Published online 10 June 2014

In vivo
In the context of this Review, *in vivo* refers to experiments carried out in living cells, regardless of whether the cells are within or outside a whole organism (sometimes referred to as *ex vivo*).

Quantitative characterization of TFBSs. The numerous technologies used in the study of TFBSs consist of two main approaches. One approach focuses on measuring the occupancy of sites along the genome, which can then be used to delineate the binding preferences of the measured TFs from a 'top-down' perspective. Using methods such as chromatin immunoprecipitation followed by microarray (ChIP–chip)[12,13], ChIP followed by high-throughput sequencing (ChIP–seq)[14,15], ChIP-exo[16] and DNase I hypersensitive site sequencing (DNase-seq)[17,18] (BOX 1), this approach can be applied *in vivo* to capture the binding events that occur along the genome in the

Box 1 | Methods for assaying protein–DNA interactions

In vitro methods

PBM. Protein binding microarray (PBM) is a high-throughput method for characterizing the *in vitro* DNA binding specificities of transcription factors (TFs). A DNA-binding protein of interest is expressed, purified and added to a double-stranded DNA (dsDNA) microarray. The microarray is then washed to remove nonspecific binding, and a fluorescent antibody is used to quantify protein binding to each probe. The probes in the microarray commonly contain all possible 10-bp potential binding sites.

HT-SELEX. High-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) can be used to characterize the *in vitro* DNA binding specificities of TFs. A dsDNA mixture (for example, a mixture containing all possible 14-bp sequences flanked by primer sites) is incubated with an immobilized DNA-binding protein. After washing, the bound oligonucleotides are recovered, amplified and used as a new set of ligands in the subsequent selection cycles. The bound dsDNA population in each cycle is subjected to high-throughput sequencing to deduce the TF binding specificities.

MITOMI. Mechanically induced trapping of molecular interactions (MITOMI) is a high-throughput method for characterizing the *in vitro* DNA binding specificities of TFs. A microfluidic device is aligned to a microarray, such that each cell contains a programmed DNA that can be bound by TFs localized to the surface. Mechanical trapping protects protein–DNA interactions, whereas unbound DNA and proteins are washed out, and the device is then scanned to quantify binding.

HiTS–FLIP. High-throughput sequencing–fluorescent ligand interaction profiling (HiTS–FLIP) can be used to characterize the *in vitro* DNA binding specificities of TFs[28]. Clusters of dsDNA are constructed on a sequencing flow cell. A fluorescently tagged protein is added and, after a wash step, binding to each cluster is quantified by visualizing fluorescence with high-throughput sequencer optics. Bound clusters are mapped to the corresponding sequences on the basis of their position on the flow cell.

DIP–chip and DIP–seq. DNA immunoprecipitation followed by microarray (DIP–chip) and DIP followed by high-throughput sequencing (DIP–seq) are high-throughput methods for identifying, on a genome-wide scale, DNA regions that are bound *in vitro* by a target protein of interest. In the DIP step, a purified DNA-binding protein is incubated with sheared genomic DNA, and protein–DNA complexes are then separated from unbound DNA using immunoprecipitation or affinity purification. Purified DNA fragments are amplified, labelled fluorescently and identified either by hybridization to a DNA microarray or by high-throughput sequencing. A similar approach to DIP–seq is termed protein–DNA binding followed by high-throughput sequencing (PB–seq).

In vivo methods

ChIP–chip and ChIP–seq. Chromatin immunoprecipitation followed by microarray (ChIP–chip) and ChIP followed by high-throughput sequencing (ChIP–seq) are high-throughput methods for identifying, on a genome-wide scale, DNA regions that are bound *in vivo* by a target protein of interest. In the ChIP step, DNA–chromatin extracts (that is, complexes of DNA and protein) are enriched by antibodies that recognize the specific DNA-binding protein of interest. The precipitated DNA is then identified through hybridization to a microarray or by high-throughput sequencing.

ORGANIC. High-resolution maps for some TFs can also be produced using occupied regions of genomes from affinity-purified naturally isolated chromatin (ORGANIC)[127]. It is a variant of ChIP–seq in which the initial steps of crosslinking and sonication that are common to ChIP applications before immunoprecipitation are replaced by digestion using micrococcal nuclease (MNase).

ChIP-exo. This is an extension of ChIP–seq that includes exonuclease trimming after immunoprecipitation to increase the resolution of the mapped binding events.

DNase-seq. DNase I hypersensitive site sequencing (DNase-seq) is a high-throughput method for identifying, on a genome-wide scale, open chromatin regions and footprints of DNA-binding proteins in these regions. Nuclei are treated with DNase I, which preferentially digests accessible DNA. Produced fragments can then be identified using microarrays or, more recently, high-throughput sequencing, in which read edges represent DNase I cleavage site. Notably, analyses of DNase I-released fragments based on their length can offer coupling of TF footprint information and nucleosome occupancy as measured on the same population in a single assay[128].

Chemical approach for nucleosome mapping. This is a high-throughput method for mapping genome-wide locations of nucleosome centres *in vivo* at single-base-pair resolution[98]. It relies on chemical modification of histones that, upon the introduction of hydrogen peroxide, results in hydroxyl radicals that cleave the DNA at sites that symmetrically flank the nucleosome centre. Cleavage patterns are then identified using high-throughput sequencing.

ATAC-seq. Assay for transposase-accessible chromatin using sequencing (ATAC-seq) is a high-throughput method for interrogating open chromatin on a genome-wide scale and capturing footprints of both nucleosomes and DNA-binding proteins[129]. Unfixed nuclei (even from a limited number of cells) are treated with a transposase and loaded *in vitro* with adaptors for high-throughput sequencing, which results in preferential integration of the adaptors in regions of accessible chromatin. Amplified DNA fragments are identified using high-throughput sequencing.

FAIRE. Formaldehyde-assisted isolation of regulatory elements (FAIRE) is a high-throughput method for identifying nucleosome-depleted regions. Cells or dissociated tissues are crosslinked briefly with formaldehyde, lysed and sonicated. Sheared chromatin is subjected to phenol–chloroform extraction, and the DNA from the aqueous phase (that is, preferentially nucleosome-depleted DNA) is purified and assayed. FAIRE-enriched fragments can be identified using microarrays (in FAIRE–chip) or high-throughput sequencing (in FAIRE–seq).

In vitro and *in vivo* methods

MNase–chip and MNase–seq. MNase digestion followed by microarray (MNase–chip) and MNase digestion followed by high-throughput sequencing (MNase–seq) are high-throughput methods for mapping nucleosomes on a genome-wide scale. Nuclei are treated with MNase, which preferentially digests accessible DNA to produce mostly mono-nucleosome-bound fragments. The resulting fragments can then be analysed using microarrays or by high-throughput sequencing. Similar to this *in vivo* application, nucleosomes can be assembled on naked genomic DNA and assayed with MNase to produce an *in vitro* genome-wide map.

Position weight matrices (PWMs). Representations for the specificity of DNA-binding proteins, in which a score is assigned to every possible base pair at each position in the binding site. A PWM score for a specific sequence is the sum of position-specific scores for each of its base pair.

measured conditions. These methods are especially beneficial for revealing how these events change across cell types and conditions, and recent applications have also attempted to use these methods to assess binding dynamics[19,20]. Although the measured binding events are not a direct readout of the TF sequence preferences but rather a reflection of the net result of several processes and effects, such techniques are used to obtain quantitative characterization of TFBSs, for example, by deriving position weight matrices (PWMs)[21] based on the frequency of occurrences of different sequences among the bound genomic regions[22].

The second approach to study TFBSs is considered to provide a more quantitative description of intrinsic sequence preferences of TFs. This approach focuses on *in vitro* affinity measurements of the chosen TFs to many short sequences, which can then be used to predict potential binding events genome wide from a

Box 2 | Identification of enhancers using massively parallel reporter assays

An ongoing challenge in the study of eukaryotic genomes is to identify enhancers and to characterize their activity. Several methods have been used to identify candidate enhancers, including computational predictions based on sequence features (for example, sequence conservation and the presence of inferred transcription factor binding site (TFBS) clusters)[58,130,131] and genome-wide measurements of transcription factor (TF) occupancy, enhancer-associated proteins (such as histone acetyltransferase p300) or chromatin features (for example, open chromatin or epigenetic markers)[35,132,133]. Formaldehyde-assisted isolation of regulatory elements (FAIRE), which isolates nucleosome-depleted DNA (BOX 1), was applied to several eukaryotic cell and tissue types to search for candidate regulatory sequences in a high-throughput manner[134].

However, although these methods generate thousands of predictions, the ability to experimentally examine them was, until recently, limited to fairly low-throughput assays (for example, luciferase reporter assays). Earlier attempts to increase the throughput were achieved by enriching for sequences that could drive fluorescent reporter expression using fluorescence-activated cell sorting (FACS), followed by subcloning and sequencing[135]. A recent study used confocal microscopy to measure the ability of ~6,500 genomic fragments to drive expression in imaginal discs[136], and in the adult and embryonic central nervous system[137,138] of *Drosophila melanogaster*.

A major increase in throughput was recently achieved by the use of DNA sequencing technologies in massively parallel reporter assays. These assays aim to both identify enhancers (when carried out on native genomic sequences) and dissect enhancer properties (when carried out on synthetic sequences) (BOX 3).

Among others, these methods allowed measurements of reporter activity of ~2,000 145-bp DNA segments that were derived from genomic sequences which showed an enhancer-associated chromatin state and that were centred on evolutionary conserved regulator motifs in 2 human cell lines[39]. An additional set of ~3,000 variants with targeted disruptions to the motifs was also measured, which generally confirmed the expected cell-specific activity of the tested regulators.

A recent method[139] eliminated the need for barcodes (which are used in mRNA sequencing-based reporter assays to assign the measured activity to each sequence variant) by placing the candidate enhancers downstream of a minimal promoter, such that active enhancers in this context are themselves transcribed and thus contained within the sequenced mRNA. This method is termed self-transcribing active regulatory region sequencing (STARR-seq) and was used to quantitatively study the activity of millions of *D. melanogaster* genomic fragments that covered >90% of the non-repetitive genome in two cell types.

A complementary method[41] termed enhancer-FACS-seq (eFS) provided a lower-throughput identification of tissue-specific enhancers in *D. melanogaster*. However, this classification was carried out for genomically integrated candidate enhancers in the context of the whole embryo. The classic microscopy examination of fluorescence is replaced with FACS-based sorting of the population of interest, followed by sequencing of candidate enhancers. This method was used to identify mesodermal enhancers among several hundred candidates, and active enhancers identified by eFS were found to be enriched with known and putative TFBSs.

Recently, site-specific genomic integration and FACS followed by sequencing (SIF–seq) was also used for enhancer identification in mammalian cells[140], specifically among >500 kb of mouse and human genomic sequences that were tested in mouse embryonic stem cells (ESCs). Additionally, fragments from specific genomic loci of interest were tested both at the initial ESC states and upon *in vitro* differentiation to cardiomyocytes and neural progenitors.

Another recently developed assay — functional identification of regulatory elements within accessible chromatin (FIREWACh)[141] — uses lentiviral-based integration rather than site-specific integration in murine ESCs and is focused primarily on reducing the search space for the identification of regulatory sequences to a relevant portion of the genome. Input sequences were produced by enriching for nucleosome-depleted regions (through incubation with restriction enzymes), thus replacing random shearing or computational pre-selection of candidates.

We note that the nature of the input sequences, and the context and conditions in which these experiments are carried out (for example, the non-endogenous location of the tested sequence, its placement within the expression cassette, its randomly fragmented nature and the use of a single, particular, minimal promoter) are likely to bare some effects on the ability of these assays to delineate endogenous enhancers and to assess their activity.

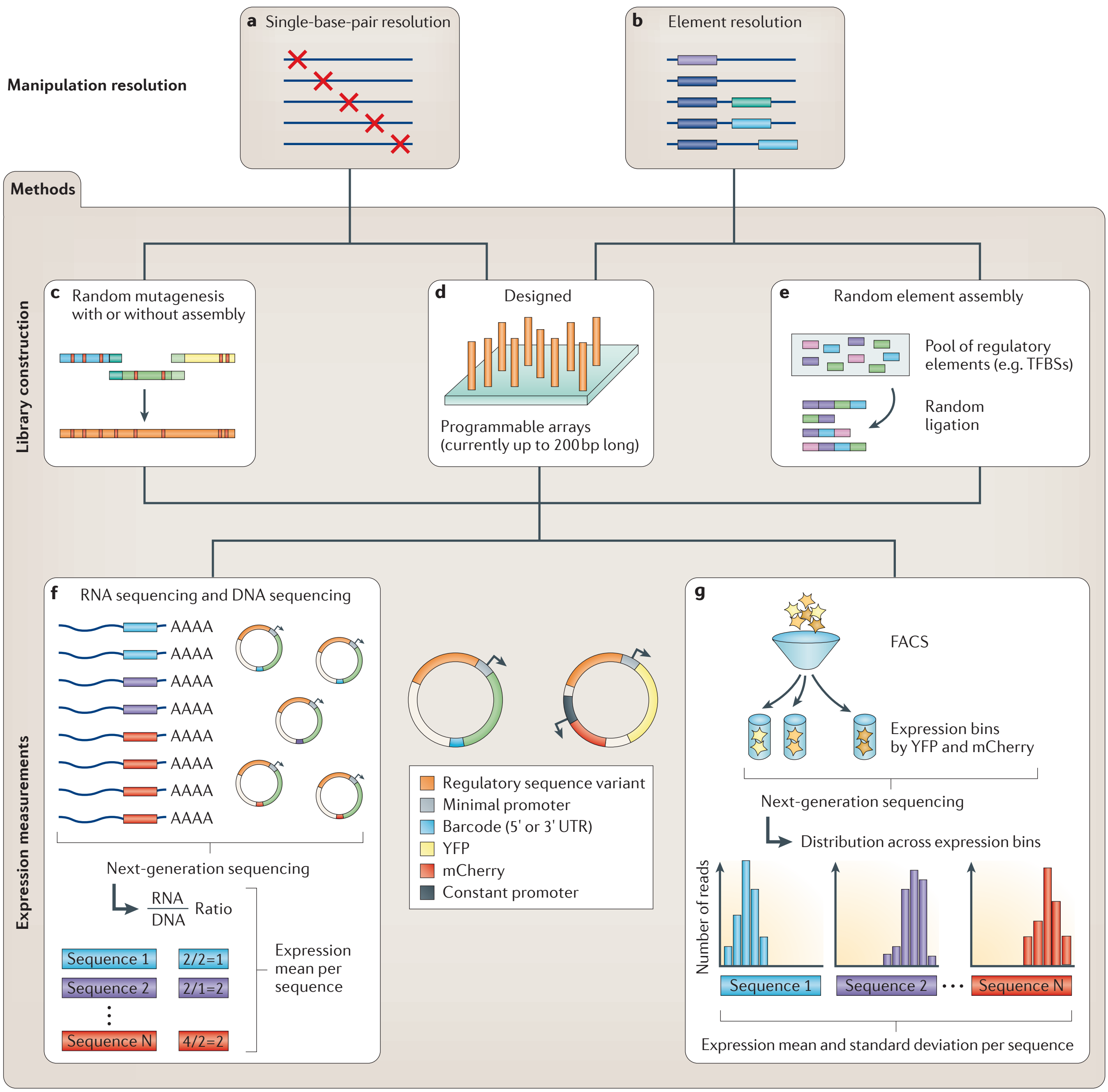
Nevertheless, the surge of high-throughput reporter assays (such as those described here and in BOX 3) provides an important step in tackling what seemed, until quite recently, to be a technological barrier that limits our understanding of the activity of regulatory sequences. These studies therefore steer future research to the challenging task of finding means to analyse and use these large-scale, rapidly accumulating data sets to gain a more mechanistic and possibly predictive understanding of sequence-encoded regulation.

Box 3 | Dissection of regulatory sequences using massively parallel reporter assays

In recent years, a surge of new methods that incorporate high-throughput sequencing into reporter assays enabled quantitative measurements of the activity of thousands of synthetically designed regulatory sequences. Some of these methods examine regulatory sequences at single-base-pair resolution (see the figure, part **a**) using extensive mutagenesis (either random (part **c**) or designed (part **d**)). Alternatively, some methods examine regulatory sequences, focusing on regulatory elements (part **b**) using random ligation (part **e**) or fully designed sequences (part **d**) that contain, for example, systematic manipulations of the composition and arrangement of transcription factor binding sites (TFBSs) . The activity of the regulatory sequences is measured by sequencing of the regulated transcripts (part **f**) or by fluorescence-activated cell sorting (FACS) according to the expression levels of the regulated reporter gene, followed by DNA sequencing to reveal the relative prevalence of each sequence variants in each of the expression bins (part **g**). Examples of applications are listed below, and their experimental trajectories are shown in parentheses.

- Studies in mouse livers (parts **a**→**c**→**f**)[76]
- Studies in newborn mouse retinas (parts **a**→**d**→**f**)[77,105]
- Studies in HEK293T (human kidney) cell line (parts **a**→**d**→**f**)[75]
- Studies in mouse livers and human HepG2 liver cells (parts **b**→**d**→**f**)[60]
- Studies in yeast cells using plasmids (parts **b**→**d**→**g**)[36] and genomic integration of constructs (parts **b**→**e**→**f**)[38]
- Studies in *Escherichia coli* (parts **a**→**c**→**g**[142]; parts **b**→**d**→**f**,**g**[143])
- *In vitro* transcription (parts **a**→**d**→**f**)[144]

UTR, untranslated region; YFP, yellow fluorescent protein.


Manipulation resolution
a Single-base-pair resolution
b Element resolution
Methods
Library construction
c Random mutagenesis with or without assembly
d Designed
Programmable arrays (currently up to 200 bp long)
e Random element assembly
Pool of regulatory elements (e.g. TFBSs)
Random ligation
Expression measurements
f RNA sequencing and DNA sequencing
AAAA
Next-generation sequencing
RNA / DNA Ratio
Sequence 1
2/2=1
Sequence 2
2/1=2
Sequence N
4/2=2
Expression mean per sequence
Regulatory sequence variant
Minimal promoter
Barcode (5′ or 3′ UTR)
YFP
mCherry
Constant promoter
g
FACS
Expression bins by YFP and mCherry
Next-generation sequencing
Distribution across expression bins
Number of reads
Sequence 1
Sequence 2
Sequence N
Expression mean and standard deviation per sequence

'bottom-up' perspective. Using methods such as protein binding microarrays (PBMs)[23], high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX)[24,25], mechanically induced trapping of molecular interactions (MITOMI)[26,27] and high-throughput sequencing–fluorescent ligand interaction profiling (HiTS–FLIP)[28] (BOX 1), the binding of hundreds of TFs from various organisms — including yeast[29,30], *Caenorhabditis elegans*[31], mice[32] and humans[25] — was thoroughly examined. The large data sets of affinity scores to each tested sequence allow a thorough examination of the various assumptions that are made on TF binding, such as the position independence assumption that underlies the popular representation of specificities with PWMs[21]. Notably, in cases in which the assay can be accurately carried out with several concentrations of a tested TF, it can be used to compute an exact dissociation constant (K_d) for each sequence[21], as shown with studies using HiTS–FLIP[28], MITOMI[26,27] and PBMs[33,34]. K_d has advantages over common binding scores, as it does not depend on the concentration of the binding molecule that was used in the experiment.

From* in vitro *specificities to* in vivo *binding and expression. The highly accurate and quantitative nature of *in vitro* measurements of binding specificity greatly advances our understanding of TF binding *in vivo*. This is shown, for example, by the identification of a strong *in vitro* characterized DNA-binding motif in most DNA segments that were found to be occupied by TFs *in vivo* in recent ChIP–seq data from the Encyclopedia of DNA Elements (ENCODE) project[15,35], and ChIP–seq peaks that lack a motif are thought to possibly indicate indirect binding events[30]. This is further supported by the fairly good agreement between the *in vitro* affinities measured for different sites of a given TF and the *in vivo* expression in yeast and human cells driven by synthetic promoters that contain these sites[36–39]. Additional support is given by the successful use of the characterized binding specificities of TFs in computational models that aim to explain complex expression patterns of native regulatory sequences[40,41].

In contrast to these studies that show the power of *in vitro*-deduced binding specificities for predicting the TFBSs bound *in vivo* and consequently the effects on gene expression, a common observation is the differential occupancy found *in vivo* for sequences that have been identified *in vitro* as preferred binders[42–46]. Such differential occupancy can stem from various mechanisms, some of which are discussed below (FIG. 1).

In some cases, the *in vitro* affinities may not accurately reflect the binding specificities *in vivo*. This can occur when the state of the binding molecule differs *in vitro* and *in vivo*, for example, when the *in vivo* binding involves interactions with cofactors that may alter the sequence preferences of the TFs[47] (FIG. 1a). In a recent study[48], the *in vitro* characterization by HT-SELEX revealed that the interaction of different *Drosophila melanogaster* Homeobox (Hox) proteins with a known DNA-binding cofactor, extradenticle (Exd), evokes binding specificities that are distinct from one another, and the differences were greater among the Hox–Exd complexes than among the Hox monomers. Notably, such alterations to binding specificities can also occur with a non-DNA-binding cofactor, as shown with the PBM-based characterization of an extended motif that is recognized by the yeast Cbf1–Met4–Met28 complex compared with the sequence specificities of Cbf1–Met4 alone[34]. In both studies, the latent specificities[49] that result from cofactor interactions contributed to the understanding of *in vivo* binding.

Notably, it is likely that even in cases in which the ranking of the binding affinities of different sequences as deduced *in vitro* is maintained *in vivo*, the quantitative differences measured *in vitro* would not accurately reflect those *in vivo*, consequently hindering our ability to understand expression differences that stem from this differential binding.

We further note that even the seemingly simpler task of predicting gene expression from *in vivo* TF binding measurements, rather than from motif occurrence directly, is still challenging. Such predictions involve deriving highly quantitative information from *in vivo* binding methods (BOX 1), but technical issues such as crosslinking and immunoprecipitation efficiencies can preclude such analyses. For example, ChIP peaks at different genomic locations and for different TFs are not readily comparable and generally do not provide an accurate quantitative measure of the fraction of cells in the population in which these regions are bound by the examined TFs. The difficulty in predicting gene expression from *in vivo* TF binding also underscores our currently limited understanding of the quantitative contribution to expression of different TF binding events[50].

One way to bridge the gap between *in vitro*-derived motif preferences, and *in vivo* TF binding and gene expression is to go beyond the isolated TFBSs and to quantitatively characterize other determinants, such as properties of regulatory architectures and the effects of sequence context.

Combining regulatory elements

TFBSs embedded into regulatory sequences such as promoters or enhancers exert their effect on gene expression by facilitating interactions between bound TFs (including both direct and indirect modes of cooperation[51]), between bound TFs and non-DNA-binding cofactors, and between bound TFs and the transcription machinery. These interactions may impose constraints on properties such as the number, location, orientation and order of TFBSs, which are frequently referred to as 'grammatical rules' of regulatory sequence architectures. It remains a major challenge to characterize these 'rules', to evaluate the degree to which they are prevalent and universal, and to assess their effect on quantitative differences in expression outcomes, transcriptional dynamics and cell-to-cell variability.

Using various approaches, many studies have addressed these issues in the past few decades. Prompted by advances in identifying regulatory sequences across different species and in annotating their embedded regulatory elements, one of such approaches is based

Dissociation constant
(K_d). The dissociation constant between two molecules (in this context, for a transcription factor and a DNA sequence). It is the ratio of the off:on rate for the formation and dissolution of the complex.

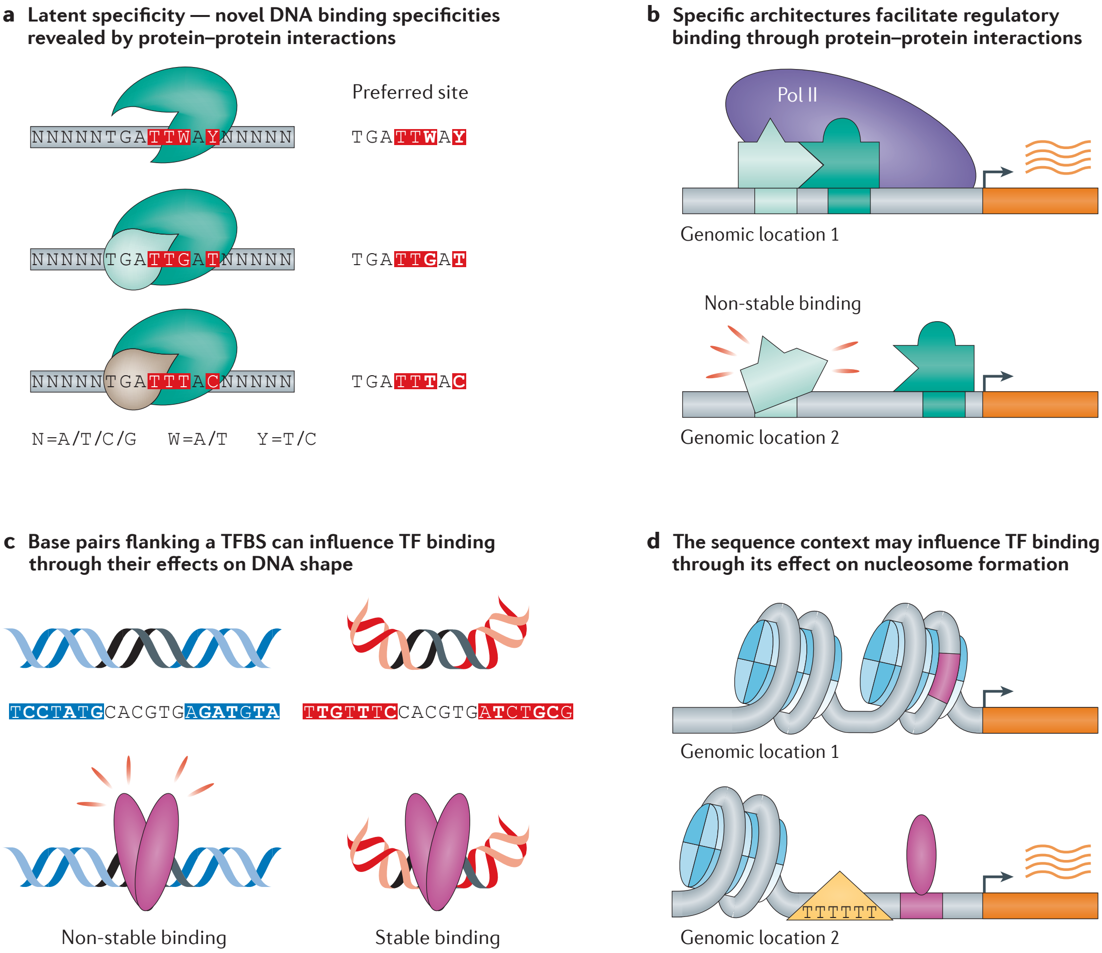

a Latent specificity — novel DNA binding specificities revealed by protein–protein interactions
Preferred site
NNNNNTGATTWAYNNNNN
TGATTWAY
NNNNNTGATTGATNNNNN
TGATTGAT
NNNNNTGATTTACNNNNN
TGATTTAC
N=A/T/C/G W=A/T Y=T/C
b Specific architectures facilitate regulatory binding through protein–protein interactions
Pol II
Genomic location 1
Non-stable binding
Genomic location 2
c Base pairs flanking a TFBS can influence TF binding through their effects on DNA shape
TCCTATGCACGTGAGATGTA
TTGTTTCCACGTGATCTGCG
Non-stable binding
Stable binding
d The sequence context may influence TF binding through its effect on nucleosome formation
Genomic location 1
TTTTTT
Genomic location 2

Figure 1 | **Mechanisms affecting transcription factor binding: beyond simple binding site specificities.** **a** | Interactions between transcription factors (TFs) or between a TF and a cofactor can result in modified DNA binding preferences, as exemplified in a recent study[48] of the binding specificities of different homeobox (Hox) protein complexes with the cofactor extradenticle (Exd) in *Drosophila melanogaster*. Accounting for these latent specificities, as opposed to relying only on the *in vitro*-deduced monomeric specificities, improves the ability to predict *in vivo* binding. **b** | As the binding affinity of a TF to the DNA and its ability to promote expression can increase as a result of protein–protein interactions with a nearby binding TF, accounting for the regulatory sequence architecture in which a predicted site is embedded (that is, the location, orientation and distance of the predicted site relative to nearby transcription factor binding sites (TFBSs)) can improve the ability to predict its probability of being bound. **c** | The base pairs flanking a TFBS can influence the binding of the TF, possibly by affecting the local DNA shape. Thus, deducing flanking preferences, either at the level of base content or, more sparsely, at the level of DNA shape features, can contribute to our understanding of differential binding to seemingly identical sites. **d** | Sequences in the vicinity of a TFBS can affect the probability of nucleosome formation, as in the case of poly(dA:dT) tracts that can alter the accessibility of the TFBS to its cognate regulator and thus indirectly affect its binding probability. Part **a** adapted with permission from REF. 48. This article was published in *Cell*, **147**, Slattery, M. *et al.*, Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins., 1270–1282, © Elsevier (2011). Part **c** adapted with permission from REF. 45. This article was published in *Cell Reports*, **3**, Gordon, R. *et al.*, Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape., 1093–1104, © Elsevier (2013).

on comparative analysis of orthologous enhancers and of enhancers that drive similar expression patterns. However, although some studies show the use of common or conserved architectural features to characterize functional organizational features[52,53], other studies advise caution in applying such inference strategies and argue that a seemingly conserved sequence property is not a clear indication of *cis*-regulatory logic[54,55]. Similarly, the intriguing observation reported in several papers that diverged regulatory regions or multiple types of architectures can produce seemingly similar expression patterns[56–59] does not in itself refute the existence or functional importance of architectural constraints but might rather allude to some degree of robustness or compensatory dynamics.

A more direct approach to address these questions is to measure the effect on expression of systematic manipulation of different properties of regulatory

Orthologous
Pertaining to loci in two species that are derived from a common ancestral locus.

architectures (see below). Whereas analyses of native regulatory sequences that differ from each other in many aspects are restricted to correlating regulatory features with the transcriptional outcome, a perturbation-based approach attempts to produce sequence variants that differ in a single parameter (for example, the presence or location of a TFBS) and thus facilitates the elucidation of the causal effect of the varied parameter. The numerous regulatory sequence manipulations carried out over the years at various loci, which commonly examined the effect in multiple tissues simultaneously, provided a wealth of insights, even though these are often qualitative in nature. Recent high-throughput technologies (BOX 3) scaled up these approaches and enable quantitative measurements of the expression driven by thousands of regulatory sequence variants in a single experiment in one or two chosen conditions or cell types.

Regulatory architecture composition. Most enhancers studied so far contain multiple regulatory elements, but the exact contribution of each regulatory element, even in a homotypic TFBS cluster, to the resulting gene expression pattern is largely unknown. Several studies showed that expression levels increase monotonically with the addition of more TFBSs and seem to saturate at a specific number of sites (FIG. 2a). This was the case for 13 TFs that were recently examined in yeast[36], as well as for 5 of 12 liver-specific TFs examined in mice and in human HepG2 cells[60]. The differences in the results obtained for these liver-specific TFs might discriminate TFs that can promote expression on their own from TFs that may require other binding partners. A further quantitative characterization of this trend in yeast was afforded by the ability to test hundreds of synthetically designed variants. Specifically, all 128 possible combinations of up to 7 TFBSs in 7 predefined locations across 2 promoter contexts were examined, which provided a direct measure of the relationship between expression and TFBS multiplicity for 2 tested TFs (by averaging across the different locations). The relationship was found to accurately follow a logistic function[36], but both the number of TFBSs at which saturation was observed and the expression value at saturation differed among TFs and sequence contexts[36,60]. This raises interesting questions about the mechanisms underlying these parameters and whether these act at the level of TF binding (for example, constraints on binding events posed by the TF concentration in the tested condition) or at the level of transcriptional activation (for example, a maximal degree of recruitment or stabilization of the transcription machinery).

Notably, as the number of binding sites for a specific TF can be a 'sensor' of that TF concentration — with higher concentrations promoting multiple binding events in promoters with several TFBSs — so can the affinity of the TFBSs (FIG. 2b). Both the presence of weak sites and the importance of constraining their affinity were shown to contribute, quantitatively and spatially, to the formation of proper expression patterns[40,61–63]. It will be interesting for future quantitative studies to characterize the aspects by which properties such as the multiplicity and affinity of TFBSs are interchangeable.

Homotypic TFBS cluster
A cluster of multiple transcription factor binding sites for the same transcription factor.

Heterotypic clustering
Clustering of multiple transcription factor binding sites for different transcription factors.

Aside from the multiplicity and affinity of TFBS clusters, it is also crucial to understand the effect of the identity of the binding TFs. The appearance of multiple TFBSs for different TFs — referred to as heterotypic clustering — leads to combinatorial regulation and allows logical-gate type of computations (such as 'AND', 'OR' and 'NOR'), the inputs of which are the presence (and, even more precisely, the concentrations) of the regulating TFs. This combinatorial capacity was shown to be important in several organisms and contexts, specifically in developmental processes that require precise readouts of morphogen gradients to form proper spatiotemporal gene expression patterns[58,64]. A recent study discussed another potential property of using multiple types of TFs: across thousands of enhancers with different combinations of 12 liver-specific TFs tested in mice and HepG2 cells, those with heterotypic TFBS clusters generally showed higher levels of expression than the corresponding homotypic clusters[60] (FIG. 2c). Unravelling the mechanism underlying this effect, which possibly involves strong cooperative interactions between the different TFs, and its relationship to the concentration of the TFs will facilitate the assessment of the generality of this observation for different TFs, cell types and conditions.

We note that, in some cases, accounting for the identity of a binding TF provides insights into the dynamics and modes of TF cooperation that occur at the examined regulatory sequence; this would be the case, for example, if one of the binding TFs is capable of binding nucleosomal DNA and increasing accessibility for other binding events — possibly through the recruitment of chromatin remodellers (such as pioneer TFs[65]) or, more generally, if it is suspected to have some potentiating role (such as the *D. melanogaster* TF Zelda)[51].

From flexible to constrained architectures. The multiplicity, identity and affinity of TFBSs are commonly referred to as the composition of a regulatory sequence. However, the degree to which composition is the key determinant of the function of a regulatory sequence is still an open question, and accumulating examples show that regulatory sequences span a range between sequences with function that largely depends on their composition and sequences with function that is also highly sensitive to the arrangement of the constituent elements (FIG. 2d).

At one end of this spectrum are regulatory sequences that can be described by the 'billboard' model[66]. This model proposes that regulatory sequences are units of information display, in which a largely flexible arrangement of regulatory elements or small modules acts in a relatively independent manner and consequently carries out a fairly additive type of computation. Enhancers that are largely insensitive to alteration in orientation, spacing or order of their constituent TFBSs may thus follow this model[54,60,67,68]. However, the activity of other enhancers may depend on the specific organization of a subset of the constituent elements (for example, two TFBSs among several), probably owing to interactions between the respective binding TFs[69,70]. Other

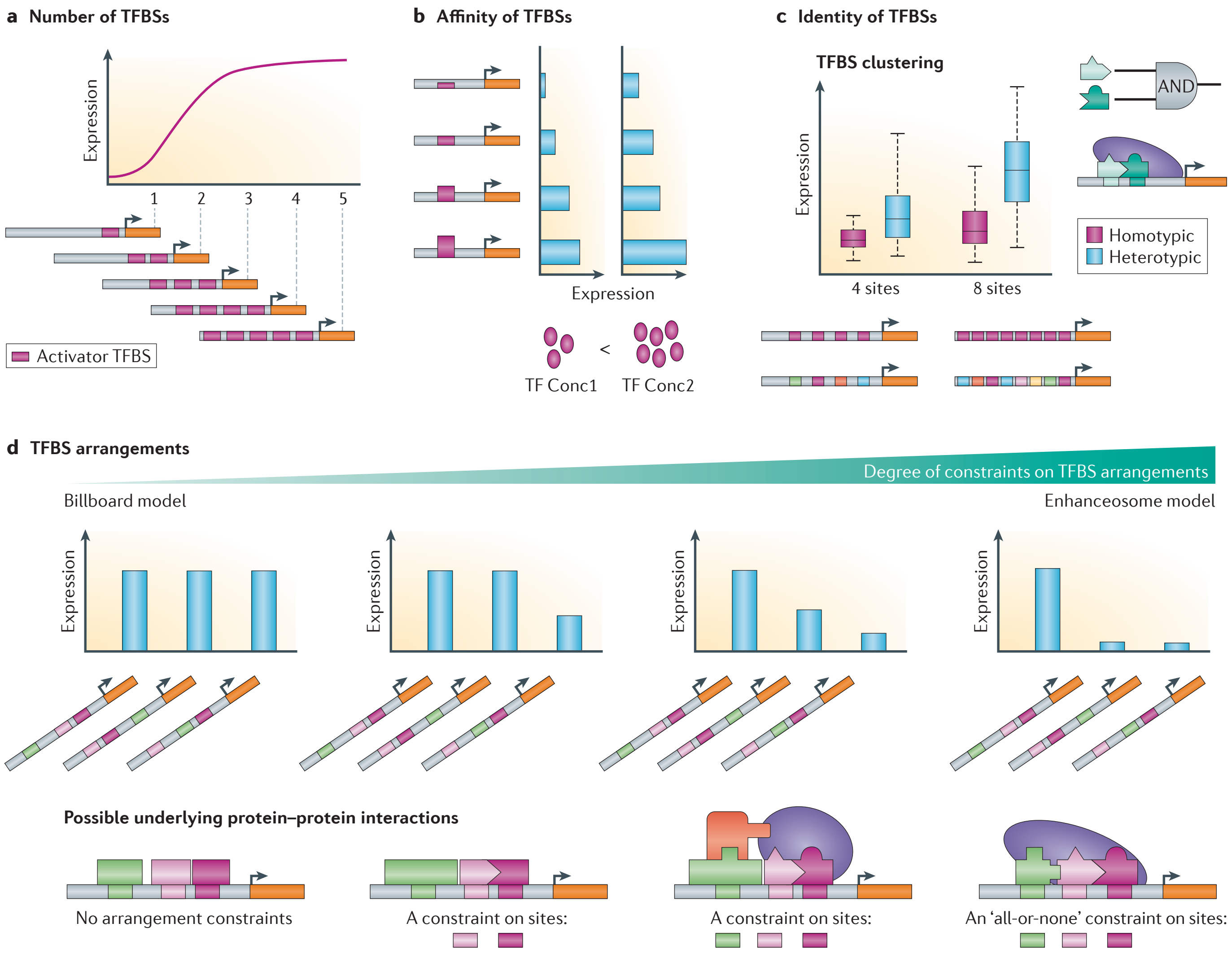


Figure 2 | **Properties of regulatory sequence architectures.** **a** | For some activators, expression was shown to increase monotonically with the addition of transcription factor binding sites (TFBSs) for that regulator and then reach saturation[36,60]. For the yeast transcription factors (TFs) Gal4 and Gcn4, this pattern perfectly matched a logistic function[36]. **b** | TFBS affinity can serve as a 'sensor' for TF concentration (conc); for example, a weak site ensures binding only above some threshold concentration. **c** | Combinatorial regulation — for example, encoding a requirement for two types of TFs (that is, an 'AND' gate) — can be attained with an architecture that includes the corresponding mixture of TFBSs. A recent study[60] suggested that such heterogeneous (that is, heterotypic) TFBS clusters can result in higher expression levels than their corresponding homogenous (that is, homotypic) TFBS clusters. **d** | Regulatory sequences vary in the degree to which their expression is sensitive to TFBS arrangements. On one end of this spectrum is the non-constrained 'billboard' model, in which expression is robust to manipulations to TFBS spacing and order. In a more constrained enhancer, expression may depend on the relative spacing of specific pairs of adjacent TFBSs. When the relative location of another TF can further enhance expression, some dependency on the TFBS order is expected. On the other end of the spectrum is the 'enhanceosome' model: when expression occurs only upon the formation of a specific complex that involves all binding TFs, a highly constrained architecture is observed, as in the case of the interferon-β enhanceosome. Part **c** from REF. 60, Nature Publishing Group.

enhancers are even more constrained, as they require particular spacing between several pairs of elements to form proper expression patterns and also show an overall sensitivity to the order of TFBSs[68,71]. Notably, constraints on specific positions and/or spacing were suggested to facilitate the interaction between TFs that bind to adjacent sites (for example, by enabling cooperativity of activators or short-range repression) or, in other cases, to inhibit such interactions (for example, by preventing promiscuous expression without the need for transcriptional repression)[62,68].

In regulatory sequences at which all DNA-binding proteins cooperate to form the activating structure, as in the case of the extensively studied interferon-β enhancer[72,73], a fixed arrangement of the underlying TFBSs might be imposed. Such a case represents the highly constrained extreme at the other end of the spectrum, which can be described by the 'enhanceosome' model[66], and these regulatory sequences are likely to drive a synergistic and hence sharp or even switch-like response.

Finally, a recent study[74] suggested another type of enhancer model termed the 'TF collective'. In this model, TF activity is determined by cooperative interactions, similarly to the enhanceosome model, yet the formation of this activating collective does not require a strict TFBS composition (that is, sites for subsets of the regulators are sufficient) or arrangement within the enhancer, which is reminiscent of the flexibility of the billboard model. Further investigations are required to better characterize this mode of operation and its prevalence.

The differences between enhancers that represent different points along this spectrum might be evident not only by direct manipulations to arrangements of regulatory elements but also in the quantitative effects on expression that results from exhaustive mutagenesis[75–77] (BOX 3). Several properties — such as the number of positions along the enhancer that significantly affect expression, the magnitude of these changes and the frequency of non-additive effects — were shown to differ between two studied enhancers, one of which was representative of the billboard model and the other of the enhanceosome model[75]. This type of classification may also underlie, to some extent, the differences in such properties reported by recent mutagenesis studies on different enhancers[76,77].

Notably, when attempting to characterize constraints on architectural properties, it is important to consider the multitude of tissues and conditions in which the manipulations to the arrangement of regulatory elements can have an effect. This can include, for example, accounting for ectopic expression[62], or expression in uninduced or stress conditions[75,78]. Several studies have elegantly shown how a specific constraint, such as the spacing between TFBSs, can seem unrestricted in one biological environment but restricted in another, or how such a constraint can have a dual role of ensuring robust expression in the desired cell type while minimizing expression in others[62,68]. These investigations highlight the notion that an observed architecture represents overlapping layers of functional information and gives a cell- or condition-specific expression readout.

As it is becoming clear that enhancers vary in the degree to and the quantitative manner by which their activity depends on architectural properties, even a very extensive and quantitative characterization of various enhancers in numerous conditions, as can now be afforded by high-throughput assays (BOX 3), does not readily lead to the emergence of 'rules' or promise to provide a principled understanding. We are thus faced with the challenge of going beyond the description of specific cases, as we attempt to elucidate whether there are principles that determine the relevance of architectural properties and constraints to the function of different enhancers under different circumstances and, if so, to characterize these principles.

Such principles can relate to the level of activation and cell-to-cell variability that is 'required' of the enhancer; the timing and duration of the enhancer activity (for example, different developmental stages[68]); the identity and families of binding regulators[79]; the type of core promoter or basal transcription machinery[80–82] with which they interact (for example, TATA versus TATA-less promoters and TFIID versus the SAGA complex); the chromatin landscape[80,81]; and other epigenetic properties of the environment of the enhancer. Studies that provide insights along these lines raise the possibility that such subdivisions of regulatory sequences in relation to their functional requirements, their genomic context and other properties can assist in elucidating distinct strategies of enhancer design and use.

Notably, in many cases, the lack of knowledge of the mechanism that underlies an observed dependency of enhancer activity on an organizational feature (FIG. 2a,c) prevents the assessment of the generality or circumstance of this dependency. Thus, a challenge for future research is to find means to provide a more mechanistic understanding of observed dependencies, in addition to an input–output description.

Beyond TFBSs: sequence context

A description of regulatory sequences that only accounts for the regulatory building blocks and their arrangements views regulatory sequences as inert strings on which functional elements are threaded. However, accumulating evidence suggests that the regulatory interactions that take place on these sequences also depend on the sequence context in which these elements are embedded. Below, we focus on two types of such contexts effects: the effects of base pairs that flank TFBSs, which may be mediated by DNA shape; and the effect of GC content, which may be mediated by nucleosome occupancy.

The effect of the flanking base pairs of TFBSs. The effect of the base pairs that flank the core TFBS was repeatedly demonstrated in the past two decades[26,28,37,83–85] and was suggested to contribute to TF binding specificity. A recent study[45] focused on two yeast basic helix–loop–helix (bHLH) TFs Cbf1 and Tye7, which were previously shown to have highly similar DNA-binding motifs (E-boxes) but seem to bind to different sets of genomic targets *in vivo*. To assess the contribution of flanking bases to the differential intrinsic preferences of these TFs, a new application of PBMs termed genomic-context PBMs (gcPBMs) was used. This allowed measurements of TF binding to variants of the core E-box motif embedded within 30-bp sequences, and the flanking bases were derived from genomic regions surrounding the TFBSs that were found to be either bound or unbound by ChIP–chip measurements. These measurements revealed the contribution of both proximal base pairs flanking the sites (which have also been discussed in previous studies[26,86]) and more distal flanking base pairs to the differential specificity of the two TFs.

One way by which flanking base pairs may influence TF binding is through their effect on the DNA structure (FIG. 1c). Indeed, an emerging view of protein–DNA recognition accounts not only for discrimination of specific base pairs through direct interactions that involve hydrogen bonds and hydrophobic contacts — the commonly acknowledged source of specificity — but also for discrimination based on various properties that are related to sequence-dependent DNA structure and deformability, which include deviations from ideal B-DNA structure (referred to as DNA shape properties)[47,87]. In support of this view, a regression model that incorporated DNA-shape features, such as groove width and rotational parameters, recapitulated the predictive power of a simple model on the basis of the occurrences of different k-mers per flanking position, with fewer explanatory variables, and elucidated distinct features of the binding of each TF examined[45].

As discussed above, many effects influence TF binding in cells, and it is thus expected that improvements in the characterization of intrinsic sequence preferences by incorporating flanking base pairs will contribute differently, depending on the TFs, to our ability to explain distinct binding patterns *in vivo*. Sequences bound *in vivo* by either Cbf1 or Tey7, as measured by ChIP–chip, mostly show a higher *in vitro* binding signal for the corresponding TF, as measured by the gcPBMs that accounted for flanking base pair preferences[45]. For another member of the bHLH family, Pho4, recent studies[37,86] showed that differences in transcription rates driven by *PHO5* promoter variants (which differ in the 1–2 base pairs flanking the E-box) were largely correlated with previous *in vitro* affinity measurements[26]. Preferences of Pho4 for particular flanking base pairs were suggested to have an important role, as they subtly differ from those of the related TF Cbf1, which was shown to compete with Pho4 for binding *in vivo*[44,86]. Notably, the degree to which the flanking base pairs contribute to TF binding can depend on the chromatin context, as shown by a generally attenuated effect of flanking base pairs for the nucleosomal Pho4-binding site compared with the exposed site in the *PHO5* promoter[37].

One specific type of flanking sequences that is suggested to have a role in TF binding are A- or T-tracts, which are thought to influence DNA bending and to induce minor groove narrowing[87]. In a recent characterization of human TF binding specificities using HT-SELEX and ChIP–seq, the core sites of many TFs were found to be flanked by 3–5 adenines or thymines, and combinations of these base pairs, which are associated with a narrow minor groove, were found to be enriched compared with combinations that are associated with a wider groove[25].

Effects of GC content mediated by nucleosome occupancy. The role of chromatin in directing TFs to their functional binding sites within the genome was discussed in numerous studies[44,88–91]. For example, such effects were suggested to underlie the considerable differences in genome-wide binding locations *in vivo* and *in vitro* (with the latter obtained using DNA immunoprecipitation followed by microarray (DIP–chip) or high-throughput sequencing (DIP–seq) assays (BOX 1)) for the TFs Lue3 in yeast[42] and Heat shock factor (Hsf) in *D. melanogaster*[43], as accounting for nucleosome occupancy or DNA accessibility in these cases improved *in vivo* data predictions relative to the *in vitro* data.

Several factors are thought to influence nucleosome occupancy and positioning[92], including the activity of chromatin remodellers and the possible role of stably bound proteins[93]. One determinant that is relevant to our discussion is the DNA sequence itself, as the affinity of histone octamers can vary for different sequences[92]. Generally, GC content is a strong predictor of nucleosome occupancy[94], although finer GC- and AT-related signals, such as those described below, have also been characterized[92]. Thus, sequences outside regulatory elements (such as TFBSs) can influence the probability of nucleosome formation, and thereby indirectly affect TF binding and gene expression[95–97], for example, through the competition between nucleosomes and TFs for DNA accessibility.

Two main sequence features are associated with nucleosome occupancy and positioning. The first consists of a ~10-bp periodic signal of AA, TT, AT or TA dinucleotides that are favoured when the DNA backbone faces inwards towards the histone core, and CC, CG, GC or GG dinucleotides when the DNA backbone faces outwards[92,98]. This feature was used, for example, to separate active enhancers from inactive ones for sets of enhancers in *Ciona intestinalis* and *D. melanogaster*, as active enhancers showed significantly lower nucleosome occupancy[99].

The second sequence feature is poly(dA:dT) tracts, which are unfavourable for nucleosome formation[92,100]. As these homopolymeric tracts, which are highly prevalent in eukaryotic promoters, were repeatedly associated with nucleosome-depleted regions, they were suggested to facilitate the accessibility of the DNA to binding TFs, thereby influencing the resulting expression (FIG. 1d). Indeed, a poly(dA:dT) tract near the TFBS of Gcn4 in the *HIS3* promoter in yeast was shown to significantly and causally affect expression[101,102]. A recent study quantitatively characterized this effect by measuring the promoter activity of a library of sequence variants that were designed for systematic manipulations to these tracts. The transcriptional effect of a poly(dA:dT) tract was found to be dependent on the composition, length and location (relative to the TFBS) of the tract, and inversely proportional to nucleosome occupancy over the TFBS. Whereas a model based only on TFBSs is inherently incapable of capturing these effects, incorporating the effects of nucleosomes and accounting for the sequence preferences of both histones and TFs yield good predictions of the transcriptional outcome based on the regulatory sequences[102].

Notably, a nucleosome-mediated effect of CG-related context features on TF binding and gene expression can be expected both when low nucleosome occupancy has a regulatory role[42–44,89,102] and when such a role is ascribed to high nucleosome occupancy[103,104].

Regression model
A model that describes the relationship between a dependent variable and one or more independent variables.

A recent study[105] provides additional support for the possible contribution of sequence context — particularly GC-related signals in the vicinity of TFBSs — specifically with regard to differential utilization of TFBSs *in vivo*. When short regions of the mouse genome bound by the cone–rod homeobox (CRX) TF (which were determined on the basis of ChIP–seq data) were used as promoters in a massively parallel reporter assay in living mouse retinas, they were found to generally drive transcriptional activation to a greater extent than unbound genomic regions with an equivalent number of CRX motifs. GC content was found to provide a strong discriminatory power between these two sets of genomic regions. Another challenge for future research is to uncover the underlying mechanisms of the suggested effect of GC-related signals on TF binding and gene expression in this and other cases, for example, those pertaining to nucleosome occupancy, DNA shape or possibly processes such as DNA methylation.

Context features as means to 'fine-tune' expression. Some manipulations to the sequence context can result in expression changes that are comparable in magnitude to those attained by direct manipulations of TFBSs[37,102]. However, in some cases, core TFBSs can only 'endure' a limited set of manipulations (as many changes result in low-affinity sites or in complete abolishment of binding), whereas context features such as properties of TFBS-flanking base pairs or of poly(dA:dT) tracts offer a wealth of possible manipulations for gradual sampling of a wide range of expression values[37,102]. Notably, the general nature of the mechanisms that are likely to underlie some of these effects, such as DNA shape- or nucleosome-mediated mechanisms, suggests that manipulations of context features can be used, both in synthetic applications and throughout evolution, as means to 'fine-tune' the expression of promoters or enhancers that are regulated by different types of TFs.

Application to GWASs and eQTL analyses

With the recent surge of available genotypic data across many individuals and the associated phenotypic data (such as disease-related states and gene expression measurements), much effort is devoted to uncovering the genomic loci or SNPs that underlie these different traits. Common approaches, such as those prevalent in GWASs, examine the frequencies of different genotypes with respect to the studied phenotypes and produce lists of significant associations. However, in many cases, such attempts have a limited ability to pinpoint the causal SNP owing to linkage disequilibrium (LD) and offer little insight into the biological mechanisms that mediate the proposed associations.

The rapidly improved characterization of genomic features could help to address these issues. Specifically, as accumulating evidence implicates DNA variation within regulatory sequences in human diseases and disorders[4–7,9], the incorporation of functionally related properties of regulatory sequences into eQTL analyses and GWASs is of great interest, as already shown in several recent studies. For example, when eQTLs and SNPs were considered with respect to DNase I hypersensitive sites (DHSs)[5,106] (BOX 1), a strong relationship emerged. Approximately 50% of the eQTLs were also found to be DNase I sensitivity QTLs (dsQTLs)[106], and disease-associated SNPs were found to concentrate within DHSs; these SNPs were further found to systematically perturb TFBSs and frequently associate with allele-specific chromatin accessibility[5]. Overall, these studies suggest the involvement of chromatin accessibility and TF binding in mediating many genotypic effects on expression and phenotypic variation. This is also supported by a recent study[107] using multiple ENCODE data sets[35] (including DNase I measurements, TF ChIP–seq data and motif information) to functionally annotate SNPs that have previously been identified in GWASs. Associated regions were found to be significantly enriched with such functional SNPs. Notably, however, in many cases the SNP with a functional role that is most strongly supported by the ENCODE data (which suggests that it is the causal SNP) is not the reported SNP in the GWAS but is instead a nearby SNP that is in LD with it. This shows the utility of accounting for functional information when searching for causal SNPs.

Alternative approaches incorporate functionally relevant annotations and genomic features not in the post-analysis steps but rather within the algorithms that attempt to identify the relevant SNPs (FIG. 3). Various annotations can be used to compute a prior probability for each SNP, which represents its likelihood of having a casual effect on gene expression. By incorporating this prior probability into a Bayesian approach to infer genotype–trait associations, a recent study[108] estimated the enrichment of causal SNPs that can provide explanations for eQTLs in different regulatory annotations, while accounting for the uncertainty in the determination of these causal SNPs. When applied to lymphoblastoid cell lines from the International HapMap Project, these identified SNPs were shown to be enriched in open chromatin regions, TFBSs and known core promoter motifs, which hints at the biological mechanisms disrupted by these variants. Another study[109] showed the use of incorporating prior probabilities based on regulatory annotations into the learning of regulatory networks.

These studies prompt the use of regulatory annotations not only to pinpoint causal SNPs but also as a means to advance the ability to predict expression from sequence. A recent study[110] attempted to address this challenging task using a different type of computational scheme — a *k*-nearest neighbours (KNN)-based approach — in which gene expression levels for some individuals can be predicted on the basis of their genotype by assessing their genotypic proximity to individuals whose genotype and expression were used to train the model. For genes with expression levels that were well predicted by this scheme — which implies the importance of *cis*-regulation to their expression — the predictions improved when genotypic proximity was estimated from a weighted contribution of the participating SNPs, in which the weights were assigned according to genomic and functional annotations. Annotations that were found to be predictive

Linkage disequilibrium
(LD). A nonrandom association of alleles at different loci (as might be observed for particular alleles at neighbouring loci that tend to be co-inherited).

DNase I sensitivity QTLs
(dsQTLs). Locations at which DNase I hypersensitive site sequencing read depth significantly correlates with the genotypes at nearby single-nucleotide polymorphisms, or insertions or deletions.

Bayesian approach
A modelling approach that uses Bayes' rule and that computes a posterior probability that a hypothesis is true using a combination of prior beliefs and observed data.

k-nearest neighbours
(KNN). A non-parametric regression method that predicts the value of a new point on the basis of the values of the *k* closest training points in the feature space.

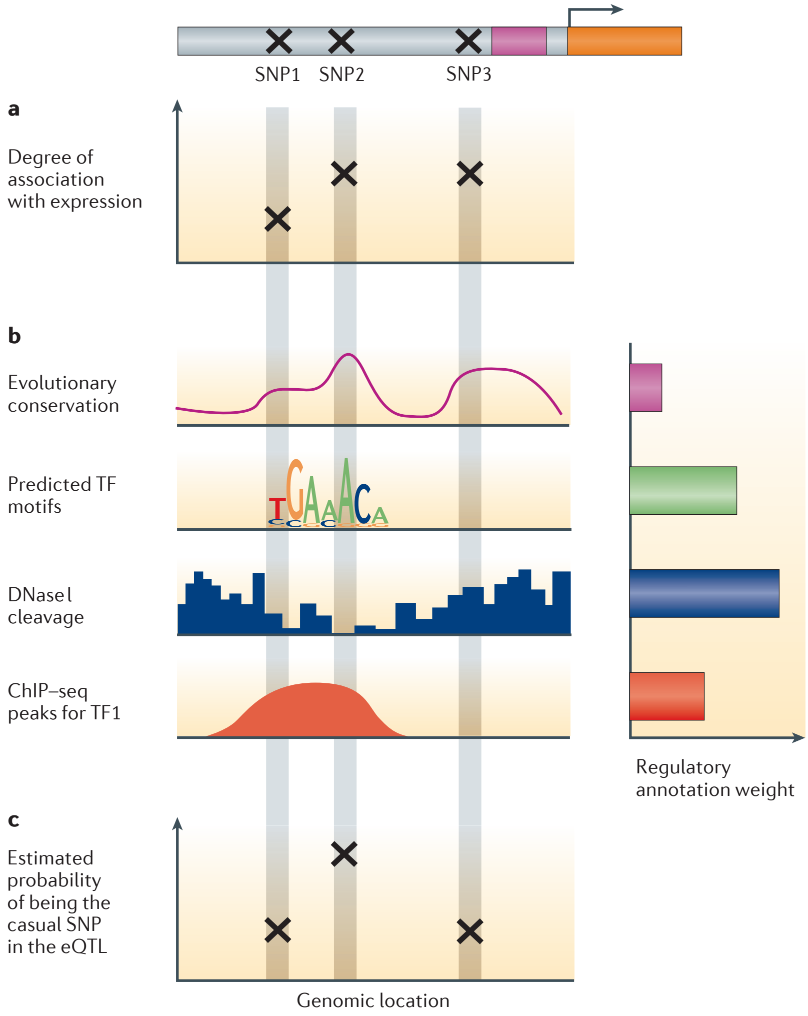

SNP1
SNP2
SNP3
a
Degree of association with expression
b
Evolutionary conservation
Predicted TF motifs
TGAAACA
DNase I cleavage
ChIP–seq peaks for TF1
Regulatory annotation weight
c
Estimated probability of being the casual SNP in the eQTL
Genomic location

Figure 3 | **A possible scheme for the incorporation of regulatory annotations to the analysis of expression quantitative trait loci.** **a** | The degree of association of different single-nucleotide polymorphisms (SNPs) to the expression level of a gene is depicted. **b** | SNPs are also examined with respect to different regulatory annotations, such that any given SNP is assigned a prior probability for being a causal site that underlies the expression quantitative trait locus (eQTL) on the basis of different regulatory annotations with varying weights (which reflect a general estimation of the prevalence of causal SNPs among each regulatory annotation that is computed on the basis of all genes). **c** | Each SNP is assigned a probability of being the causal site by combining the information from parts **a** and **b**, such that even if two SNPs had a similar degree of association with gene expression, one could receive a higher probability based on its regulatory annotations. ChIP–seq, chromatin immunoprecipitation followed by high-throughput sequencing; TF, transcription factor. Adapted with permission from REF. 108, BioMed Central.

include proximity to transcription start sites (TSSs), high regional GC content, and presence in microRNAs and TFBSs.

As our understanding of sequence-encoded regulation advances, a wider range of regulatory annotation can be integrated into such studies; these annotations could include accurate description of TFBSs (with differential quantitative effects to different disruptions within sites), of properties of regulatory architectures (for example, by assessing TFBS functionality on the basis of the distance from another TFBS that is required to form an activating complex) and of sequence context features (for example, by accounting for perturbation of base pairs flanking a TFBS that is predicted to alter DNA shape). Thus, the improved characterization of regulatory sequences, together with approaches that incorporate this knowledge into GWASs and eQTL studies, offers great promise to advance the understanding of the mechanisms that underlie the genotype–phenotype transition.

Concluding remarks

Recent studies have substantially improved our ability to 'read' regulatory information encoded within non-coding DNA sequences. Although consisting of only four 'letters', the regulatory DNA 'language' is complex and rich, and different combinations of 'words' have different meanings that, in some cases, depend on their arrangement and/or appearance within a specific context.

An additional layer of complexity arises from the fact that the same regulatory sequence may have different readouts in different *trans*-environments (for example, under different concentrations of the regulators). Further research is required to obtain a better quantitative characterization of this dependency and a principled understanding, if possible, of the degree to which different regulatory sequences lend themselves to different cellular interpretations.

Notably, in this Review and consistent with many other studies, we implicitly adopt the view that transcriptional regulation is a two-step process, in which sequence first determines binding configurations that in turn regulate gene expression. Although beneficial as a modelling approach[111,112], the extent to which this view indeed captures *in vivo* dynamics is unclear. Can expression indeed be perceived as independent of the regulatory sequence given a correct account of the 'intermediate layer' of binding? What properties should be taken into account in order to allow this independence? How do studies indicating that binding is not synonymous with functionality[51] fit into this view?

In an analogous manner to our improved understanding of enhancers, our knowledge of core promoters is gradually improving with the characterization of regulatory elements[113,114], properties of their arrangements[113] and the effects of sequence context[115] (which influences, for example, RNA polymerase II (Pol II) scanning and pausing). Studies pertaining to long-range regulatory interactions and specifically enhancer–core promoter communication (BOX 4) will pave the way to a more comprehensive view of the combined activity of these regulatory sequences. Additionally, our understanding of DNA sequence features that can influence gene expression control extends beyond those related only to transcriptional regulation and also includes sequence properties in 3′ and 5′ untranslated regions, and in coding regions (for example, properties that affect mRNA stability[116,117], splicing[118] and translation efficiency[119]). Recent studies (such as those using genome-wide measurements of transcription and Pol II recruitment) also advance our

Box 4 | Combining regulatory sequences to confer long-range regulation

Unlike regulatory sequences in yeast that are generally adjacent to their target gene, enhancers in flies and mammals are located at varying distances from their regulated gene and form long-range enhancer–promoter interactions[51,145]. To understanding how enhancers promote gene expression, it is thus necessary to replace the conventional view of the genome as a linear entity with a structurally complex and dynamic view. Substantial advances were recently made both by microscopy-based studies and by the complementary development of chromosome conformation capture (3C) methods (reviewed in REFS 146,147). These methods facilitated the characterization of genome organization at different genomic scales[146], including the recent identification of subchromosomal spatial domains (which range from ~100 kb to megabases in size) known as topologically associating domains (TADs)[146,147]. TADs are characterized by higher frequency of intra-domain long-range associations than inter-domain associations. They are conserved to a large extent among cell types and between mice and humans[146,147], and are also less variable than larger-scale structures at the single-cell level[148]. Notably, these methods also indicate that although enhancers do not necessarily interact with their nearest promoters[149], the majority of interactions seem to occur within TADs[146,147]. This suggests that the identification of these domains can provide a largely cell-type-invariant repertoire of possible regulatory interactions, and the dynamic nature of internal interaction represents differential, cell type-specific use of this repertoire of interactions.

These methods therefore substantially advance our knowledge of 'who goes with whom', yet understanding what regulates these interactions and realizing their implications on gene expression remain largely open questions. Several studies discussed possible architectural roles of proteins that interact with the DNA, including CCCTC-binding factor (CTCF, for which binding sites are enriched in TAD boundaries, although they are also found within TADs)[147], cohesin, Mediator[150] and specific transcription factors that are implicated in enhancer–promoter interactions (the knockdown of which was found to disrupt spatial organization)[146]. However, their exact involvement in gene regulation is still unclear. Generally, little is known about regulatory sequence features (for example, motif occurrence) that pertain to their capacity to form different spatial organizations.

An additional layer of complexity stems from the observation that many promoters are regulated by multiple enhancers, which raises questions about how regulatory information is distributed and integrated. In some cases, it seems that proper gene expression requires information from multiple regulating enhancers, and a specific enhancer was shown to contribute to the formation of a 'sharper' border of a developmentally related pattern or to an increased level of expression[151]. In other cases, the enhancers seem to drive similar expression patterns, and the enhancer that is located further away (which has been commonly discovered in unexpected locations such as introns of neighbouring genes) is referred to as the 'shadow enhancer' (REFS 78,151). Several studies suggested that such enhancers may provide robustness to expression — a notion that was supported by the loss of their seemingly redundant nature in suboptimal conditions (for example, at extreme temperatures or in a compromised genetic background). One hypothesized mechanism proposes that reliability of gene expression is ensured in these cases by increasing the probability of desired enhancer–promoter interactions, thus buffering a reduction in interaction efficiency of any single enhancer[78,147]. Although this represents one example of the logic that might underlie distribution of regulatory information among multiple enhancers, another possible benefit may stem from the separation of binding activators and repressors to prevent the formation of undesired short-range interactions between regulators[151].

A goal for future research is thus to advance our understanding of the relationship between sequence features — such as the composition, arrangement and sequence context of regulatory elements within enhancers, properties of core promoters and the distance between them[145] — and both the formation of long-range enhancer–promoter interactions and the nature of information processing carried out at these regulatory sequences.

understanding of transcriptional events that result in non-coding transcripts, including the prevalent enhancer-derived transcripts termed enhancer RNAs (eRNAs)[120,121]. As the functional role of eRNAs is still under investigation[120,121], it is not clear whether they impose additional constraints on enhancer sequences and how transcript levels quantitatively depend on the properties of the enhancer. Nevertheless, the improved characterization of eRNAs and the ability to to readily identify them in a high-throughput manner already facilitated the use of these RNAs for identifying enhancers[120,122], which provides a complementary approach to massively parallel reporter assays (BOX 2). Future challenges include the integration of these mechanisms of regulation and additional aspects of epigenetics regulation to provide a complete view of both the formation and the maintenance of regulatory programmes.

Importantly, this Review focused primarily on the effects of regulatory DNA on TF binding and gene expression as measured at the cell population level. A great challenge for coming years, which is already well under way, is to capture regulatory dynamics at the single-cell level, thus providing a deeper mechanistic understanding and revealing the degree of cell-to-cell variability and the exact distributions that underlie the population measure.

Finally, advances in our understanding of the means by which regulatory sequences affect various aspects of the transcriptional output (that is, gene expression levels, cell-to-cell variability and induction dynamics) pave the way for the design of regulatory sequences or circuits that produce a desired outcome, which has applications in both synthetic biology and gene therapy. Such applications are continuously emerging and are further facilitated by the recent development of genome editing tools[123–126]. Thus, although many open questions remain, we are gradually advancing in our ability to read and 'write' individual genomes.

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
2. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**, R754–R763 (2010).
3. Williamson, I., Hill, R. E. & Bickmore, W. A. Enhancers: from developmental genetics to the genetics of common human disease. *Dev. Cell* **21**, 17–19 (2011).
4. Dickel, D. E., Visel, A. & Pennacchio, L. A. Functional anatomy of distant-acting mammalian enhancers. *Phil. Trans. R. Soc. B* **368**, 20120359 (2013).
5. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
6. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
7. Sakabe, N. J., Savic, D. & Nobrega, M. A. Transcriptional enhancers in development and disease. *Genome Biol.* **13**, 238 (2012).
8. Cowper-Sal lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genet.* **44**, 1191–1198 (2012).
9. Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L. A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer Res.* **73**, 4180–4184 (2013).
10. Struhl, K. Yeast transcriptional regulatory mechanisms. *Annu. Rev. Genet.* **29**, 651–674 (1995).
11. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997).
12. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
13. Venters, B. J. *et al.* A comprehensive genomic binding map of gene and chromatin regulatory proteins in *Saccharomyces*. *Mol. Cell* **41**, 480–492 (2011).
14. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
15. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* **22**, 1723–1734 (2012).
16. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
This paper presents ChIP-exo, which is an extension of the ChIP–seq protocol. This method substantially improves the accuracy of identifying genomic locations of DNA binding events by using an exonuclease to trim immunoprecipitated DNA to a precise distance from the crosslinking site. It was applied to several yeast TFs and to human CCCTC-binding factor (CTCF). In later studies, it was also applied to yeast pre-initiation complexes and to human initiation factors, which provided mechanistic insights into transcription initiation.
17. Hesselberth, J. R. *et al.* Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).
18. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
19. Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G. & Lieb, J. D. Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**, 251–255 (2012).
20. Poorey, K. *et al.* Measuring chromatin interaction dynamics on the second time scale at single-copy genes. *Science* **342**, 369–372 (2013).
21. Stormo, G. D. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nature Rev. Genet.* **11**, 751–760 (2010).
22. MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae. BMC Bioinformatics* **7**, 113 (2006).
23. Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protoc.* **4**, 393–411 (2009).
24. Zhao, Y., Granas, D. & Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS Comput. Biol.* **5**, e1000590 (2009).
25. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
26. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
27. Fordyce, P. M. *et al. De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotech.* **28**, 970–975 (2010).
28. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotech.* **29**, 659–664 (2011).
29. Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* **32**, 878–887 (2008).
30. Zhu, C. *et al.* High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* **19**, 556–566 (2009).
31. Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**, 314–327 (2009).
32. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
33. Jones, R. B., Gordus, A., Krall, J. A. & MacBeath, G. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* **439**, 168–174 (2006).
34. Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. & Bulyk, M. L. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 555 (2011).
35. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
36. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotech.* **30**, 521–530 (2012).
37. Rajkumar, A. S., Denervaud, N. & Maerkl, S. J. Mapping the fine structure of a eukaryotic promoter input–output function. *Nature Genet.* **45**, 1207–1215 (2013).
This study measures the activity of ~200 variants of the *PHO5* promoter in yeast that differ in the binding site for the regulating TF Pho4. Temporal promoter activity measurements throughout induction were obtained with a microfluidic-based platform. Previously characterized *in vitro* affinities were found to be highly predictive of the activity of the corresponding promoter variants *in vivo*. Subtle tuning of promoter activity could be achieved by manipulating the base pairs flanking the TFBS core.
38. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the *in vivo* effects of binding site variants. *Genome Res.* **23**, 1908–1915 (2013).
39. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
40. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
41. Gisselbrecht, S. S. *et al.* Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. *Nature Methods* **10**, 774–780 (2013).
42. Liu, X., Lee, C. K., Granek, J. A., Clarke, N. D. & Lieb, J. D. Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
43. Guertin, M. J., Martins, A. L., Siepel, A. & Lis, J. T. Accurate prediction of inducible transcription factor binding intensities *in vivo*. *PLoS Genet.* **8**, e1002610 (2012).
44. Zhou, X. & O'Shea, E. K. Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol. Cell* **42**, 826–836 (2011).
This paper sets out to bridge the gap between the frequent occurrences of the TF Pho4 motif along the genome and its binding pattern *in vivo*. It suggests that several mechanisms are at play. Nucleosome occupancy seems to restrict Pho4 binding, which is further tuned by competition with Cbf1 — another TF that has similar sequence preferences. A cooperative interaction between Pho4 and a nearby binding Pho2 is further required to activate transcription.
45. Gordan, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* **3**, 1093–1104 (2013).
This study uses an extension of PBM to characterize the binding specificities of two E-box-binding TFs — Cbf1 and Tye7 — for putative binding sites in their genomic context. A differential specificity based on the base pairs flanking the core motif is characterized, and a computational model suggests that such specificity is mediated by distinct preferences for three-dimensional DNA shape properties.
46. Bresnick, E. H., Katsumura, K. R., Lee, H. Y., Johnson, K. D. & Perkins, A. S. Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Res.* **40**, 5819–5831 (2012).
47. Siggers, T. & Gordan, R. Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.* **42**, 2099–2111 (2014).
48. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270–1282 (2011).
49. Lelli, K. M., Slattery, M. & Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* **46**, 43–68 (2012).
50. Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21**, 611–626 (2011).
51. Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nature Rev. Genet.* **13**, 613–626 (2012).
52. Papatsenko, D., Goltsev, Y. & Levine, M. Organization of developmental enhancers in the *Drosophila* embryo. *Nucleic Acids Res.* **37**, 5665–5677 (2009).
53. Erives, A. & Levine, M. Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **101**, 3851–3856 (2004).
54. Rastegar, S. *et al.* The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev. Biol.* **318**, 366–377 (2008).
55. Lusk, R. W. & Eisen, M. B. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. *PLoS Genet.* **6**, e1000829 (2010).
56. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106 (2008).
57. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66–74 (2010).
58. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462**, 65–70 (2009).
59. Brown, C. D., Johnson, D. S. & Sidow, A. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317**, 1557–1560 (2007).
60. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genet.* **45**, 1021–1028 (2013).
61. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
62. Evans, N. C., Swanson, C. I. & Barolo, S. Sparkling insights into enhancer structure, function, and evolution. *Curr. Top. Dev. Biol.* **98**, 97–120 (2012).
This review focuses on the sparkling eye enhancer of the *D. melanogaster Pax2* (also known as *sv*) gene. It discusses various analyses, including the examination of sparkling orthologues and the expression measurements in several cell types of the effects of different manipulations to the composition and arrangement of TFBSs. These analyses reveal a complex combinatorial code that is densely encoded in the enhancer and several highly constrained architectural properties to ensure proper cell-specific expression.
63. Parker, D. S., White, M. A., Ramos, A. I., Cohen, B. A. & Barolo, S. The *cis*-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Sci Signal* **4**, ra38 (2011).

64. Rogers, K. W. & Schier, A. F. Morphogen gradients: from generation to interpretation. *Annu. Rev. Cell Dev. Biol.* **27**, 377–407 (2011).
65. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
66. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.* **94**, 890–898 (2005).
67. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 (1996).
68. Liu, F. & Posakony, J. W. Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules. *PLoS Genet.* **8**, e1002796 (2012).
This study examines the contribution of architectural properties of two Notch-regulated enhancers to their spatially distinct activities. Although one enhancer is resistant, to a large extent, to manipulations in the arrangement of its constituent TFBSs, the other enhancer is highly sensitive. The authors discuss how this differential reliance on architectural properties may be linked to the different developmental stages and contexts in which these enhancers function.
69. Senger, K. *et al.* Immunity regulatory DNAs share common organizational features in *Drosophila. Mol. Cell* **13**, 19–32 (2004).
70. Crocker, J., Tamori, Y. & Erives, A. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol.* **6**, e263 (2008).
71. Swanson, C. I., Evans, N. C. & Barolo, S. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev. Cell* **18**, 359–370 (2010).
72. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-β enhanceosome. *Cell* **129**, 1111–1123 (2007).
73. Thanos, D. & Maniatis, T. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).
74. Junion, G. *et al.* A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473–486 (2012).
75. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotech.* **30**, 271–277 (2012).
76. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo. Nature Biotech.* **30**, 265–270 (2012).
77. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc. Natl Acad. Sci. USA* **109**, 19498–19503 (2012).
78. Lagha, M., Bothma, J. P. & Levine, M. Mechanisms of transcriptional precision in animal development. *Trends Genet.* **28**, 409–416 (2012).
79. Guo, Y., Mahony, S. & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **8**, e1002638 (2012).
80. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084–1091 (2008).
This analysis of yeast promoters suggests two typical promoter structures that differ in their nucleosome positions, TFBS composition and location, expression variation and transcriptional plasticity (which is a measure of the degree by which gene expression is modulated across conditions); it contributes to our understanding of how different promoter architectures and dynamics may be used to attain different functional properties of expression.
81. Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
82. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
83. Leonard, D. A., Rajaram, N. & Kerppola, T. K. Structural basis of DNA bending and oriented heterodimer binding by the basic leucine zipper domains of Fos and Jun. *Proc. Natl Acad. Sci. USA* **94**, 4913–4918 (1997).
84. Morin, B., Nichols, L. A. & Holland, L. J. Flanking sequence composition differentially affects the binding and functional characteristics of glucocorticoid receptor homo- and heterodimers. *Biochemistry* **45**, 7299–7306 (2006).
85. Nagaoka, M., Shiraishi, Y. & Sugiura, Y. Selected base sequence outside the target binding site of zinc finger protein Sp1. *Nucleic Acids Res.* **29**, 4920–4929 (2001).
86. Aow, J. S. *et al.* Differential binding of the related transcription factors Pho4 and Cbf1 can tune the sensitivity of promoters to different levels of an induction signal. *Nucleic Acids Res.* **41**, 4877–4887 (2013).
87. Rohs, R. *et al.* Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
88. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).
89. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
90. Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genet.* **7**, e1001290 (2011).
91. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genet.* **43**, 264–268 (2011).
92. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nature Struct. Mol. Biol.* **20**, 267–273 (2013).
93. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4**, e1000138 (2008).
94. Tillo, D. & Hughes, T. R. G + C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**, 442 (2009).
95. Narlikar, L., Gordan, R. & Hartemink, A. J. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.* **3**, e215 (2007).
96. Raveh-Sadka, T., Levo, M. & Segal, E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.* **19**, 1480–1496 (2009).
97. Wasson, T. & Hartemink, A. J. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.* **19**, 2101–2112 (2009).
98. Brogaard, K., Xi, L., Wang, J. P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496–501 (2012).
This paper presents a chemical-based approach to map nucleosome positions genome wide at a single-base-pair resolution. This method reveals overlapping positions within the population and allows a high-resolution examination of nucleosome positions relative to sequence and genomic features such as TSS, TFBSs and Pol II pause sites.
99. Khoueiry, P. *et al.* A *cis*-regulatory signature in ascidians and flies, independent of transcription factor binding sites. *Curr. Biol.* **20**, 792–802 (2010).
100. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).
101. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).
102. Raveh-Sadka, T. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature Genet.* **44**, 743–750 (2012).
103. Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5**, e9129 (2010).
104. Ballare, C. *et al.* Nucleosome-driven transcription factor binding and gene regulation. *Mol. Cell* **49**, 67–79 (2013).
105. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel *in vivo* enhancer assay reveals that highly local features determine the *cis*-regulatory function of ChIP–seq peaks. *Proc. Natl Acad. Sci. USA* **110**, 11952–11957 (2013).
106. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
107. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
108. Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
109. Lee, S. I. *et al.* Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* **5**, e1000358 (2009).
110. Manor, O. & Segal, E. Robust prediction of expression differences among human individuals using only genotype information. *PLoS Genet.* **9**, e1003396 (2013).
111. Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15**, 116–124 (2005).
112. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature Rev. Genet.* **10**, 443–456 (2009).
113. Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 40–51 (2012).
114. Mogno, I., Vallania, F., Mitra, R. D. & Cohen, B. A. TATA is a modular component of synthetic promoters. *Genome Res.* **20**, 1391–1397 (2010).
115. Lubliner, S., Keren, L. & Segal, E. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.* **41**, 5569–5581 (2013).
116. Schoenberg, D. R. & Maquat, L. E. Regulation of cytoplasmic mRNA decay. *Nature Rev. Genet.* **13**, 246–259 (2012).
117. Burgess, D. J. Global analyses of determinants of RNA decay. *Nature Rev. Genet.* http://dx.doi.org/10.1038/nrg3710 (2014).
118. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Rev. Mol. Cell Biol.* **14**, 153–165 (2013).
119. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Rev. Genet.* **15**, 205–213 (2014).
120. Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* **46**, 1–19 (2012).
121. Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* **39**, 170–182 (2014).
122. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
123. Lohmueller, J. J., Armel, T. Z. & Silver, P. A. A tunable zinc finger-based framework for Boolean logic computation in mammalian cells. *Nucleic Acids Res.* **40**, 5180–5187 (2012).
124. Teo, W. S. & Chang, M. W. Development and characterization of AND-gate dynamic controllers with a modular synthetic *GAL1* core promoter in *Saccharomyces cerevisiae. Biotechnol. Bioeng.* **111**, 144–151 (2013).
125. Perez-Pinera, P. *et al.* Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nature Methods* **10**, 239–242 (2013).
126. Gaj, T., Gersbach, C. A. & Barbas, C. F. 3rd. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).
127. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nature Methods* **11**, 203–209 (2014).
128. Vierstra, J., Wang, H., John, S., Sandstrom, R. & Stamatoyannopoulos, J. A. Coupling transcription factor occupancy to nucleosome architecture with DNase–FLASH. *Nature Methods* **11**, 66–72 (2014).
129. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).
130. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **3**, 30 (2002).
131. Berman, B. P. *et al.* 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**, 757–762 (2002).
132. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

133. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
134. Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature Protoc.* **7**, 256–267 (2012).
135. Khambata-Ford, S. *et al.* Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Res.* **13**, 1765–1774 (2003).
136. Jory, A. *et al.* A survey of 6,300 genomic fragments for *cis*-regulatory activity in the imaginal discs of *Drosophila melanogaster. Cell Rep.* **2**, 1014–1024 (2012).
137. Jenett, A. *et al.* A GAL4-driver line resource for *Drosophila* neurobiology. *Cell Rep.* **2**, 991–1001 (2012).
138. Manning, L. *et al.* A resource for manipulating gene expression and analyzing *cis*-regulatory modules in the *Drosophila* CNS. *Cell Rep.* **2**, 1002–1013 (2012).
139. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
140. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nature Methods* **11**, 566–571 (2014).
141. Murtha, M. *et al.* FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nature Methods* **11**, 559–565 (2014).
142. Kinney, J. B., Murugan, A., Callan, C. G. Jr & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci. USA* **107**, 9158–9163 (2010).
143. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in *Escherichia coli. Proc. Natl Acad. Sci. USA* **110**, 14024–14029 (2013).
144. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotech.* **27**, 1173–1175 (2009).
145. Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
146. Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49**, 773–782 (2013).
147. Smallwood, A. & Ren, B. Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.* **25**, 387–394 (2013).
148. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
149. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
150. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
151. Barolo, S. Shadow enhancers: frequently asked questions about distributed *cis*-regulatory information and enhancer redundancy. *Bioessays* **34**, 135–141 (2012).

Acknowledgements

This work was supported by grants from the European Research Council and the US National Institutes of Health to E.S. M.L. thanks the Azrieli Foundation for the award of an Azrieli Fellowship.

Competing interests statement

The authors declare no competing interests.