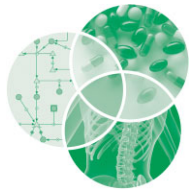
Advanced Review



Advances in analysis of transcriptional regulatory networks

Tae-Min Kim and Peter J. Park*

A transcriptional regulatory network represents a molecular framework in which developmental or environmental cues are transformed into differential expression of genes. Transcriptional regulation is mediated by the combinatorial interplay between *cis*-regulatory DNA elements and *trans*-acting transcription factors, and is perhaps the most important mechanism for controlling gene expression. Recent innovations, most notably the method for detecting protein–DNA interactions genome-wide, can help provide a comprehensive catalog of *cis*-regulatory elements and their interaction with given *trans*-acting factors in a given condition. A transcriptional regulatory network that integrates such information can lead to a systems-level understanding of regulatory mechanisms. In this review, we will highlight the key aspects of current knowledge on eukaryotic transcriptional regulation, especially on known transcription factors and their interacting regulatory elements. Then we will review some recent technical advances for genome-wide mapping of DNA–protein interactions based on high-throughput sequencing. Finally, we will discuss the types of biological insights that can be obtained from a network-level understanding of transcription regulation as well as future challenges in the field. © 2010 John Wiley & Sons, Inc. *WIREs Syst Biol Med* 2011 3 21–35 DOI: 10.1002/wsbm.105

INTRODUCTION

Cellular programs for development, responses to environmental cues, homeostasis, and all other essential processes in higher eukaryotes are mediated by a set of complex yet highly coordinated regulatory mechanisms. How is such coordination achieved in a temporally and spatially regulated manner? What genes are involved in specific cellular programs and what are the relationships (e.g., regulator vs regulated genes) among them? To gain insights on these fundamental biological questions, many biochemical and genetic studies have focused on identifying the gene regulatory networks underlying these mechanisms.

While gene expression is controlled at many levels, transcriptional regulation is perhaps the most extensively investigated mechanism.[1,2] Transcriptional control is executed by the interaction between *trans*-acting transcription factors as regulators and *cis*-regulatory sequence elements of their target genes. Individual genes generally have their own sets of *cis*-regulatory elements, which provide a multitude of binding sites for different transcription factors. Thus, the lack of a comprehensive catalog of these transcriptional regulatory elements (*cis*-elements and *trans*-factors) and the characterization of their interactions have been major issues in gene regulation.

As the role of transcriptional control has been characterized for critical cellular functions (e.g., the cell cycle) and developmental programs (e.g., body patterning), its clinical relevance has also become clear. Certain disease phenotypes in humans can arise from dysfunction of known transcription factors or through mutations of *cis*-regulatory elements. It has been reported that certain mutations of *cis*-regulatory elements can have significant phenotypic consequences, often with more profound effects than mutations in the coding regions.[3] Diseases that arise due to inappropriate activities of transcription factors are summarized in recent reviews.[4,5] Recently, it was estimated that more than 150 transcription factors are associated with ~300 human diseases.[6] A more

*Correspondence to: peter_park@harvard.edu

Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA

DOI: 10.1002/wsbm.105

 © 2010 John Wiley & Sons, Inc.

complete understanding of the transcriptional networks can thus have significant clinical applications.

Recent progress in high-throughput sequencing technology is having a tremendous impact on our understanding of the transcriptional regulatory mechanisms. In particular, chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) allows precise estimation of the locations where the protein of interest is bound. Coupled with RNA sequencing, which measures the amount of mRNA in the cell, and other high-throughput methods, these data are likely to lead us to new insights on how a transcriptional regulatory network is encoded and executed. In this review, we will first summarize previous approaches for studying transcriptional control including genome-wide motif discovery methods. Then, we will review recent technical advancements in genome-wide mapping of DNA–protein interactions with focus on ChIP-seq. Finally, we will review the network-based perspective of transcriptional regulation and how multiple sources of information can be integrated into the construction of transcriptional regulatory networks.

TRANSCRIPTION FACTORS AND DNA ELEMENTS

Transcription Factors

Transcription control is directed primarily by two elements: DNA-binding proteins (in particular, transcription factors) and DNA sequences that facilitate the binding of these proteins (*cis*-regulatory elements). Transcription factors are typically composed of two functional domains: a DNA-binding domain and a transcription regulation domain.[7] The DNA-binding domain recognizes a specific DNA sequence and affixes the transcription factor to the DNA. The transcription regulation domains can bind to other transcriptional regulatory proteins and exert regulatory functions. There are currently more than 100 known binding domains that have been used to predict potential transcription factors.[8] In the human genome, about 10% of gene products are predicted to have DNA-binding properties and are therefore considered candidate transcription factors. A recent work to curate this list of candidates has yielded a catalog of roughly 1400 transcription factors.[6] Similar studies have estimated the number of transcription factors in other model organisms: ~200 factors (3% of coding genes) in *Saccharomyces cerevisiae*[9] and ~900 factors (5% of coding genes) in the nematode *Caenorhabditis elegans*.[10] This suggests that the number of candidate transcription factors per gene may increase with genome size and organism complexity[11]; however, complex transcriptional control of higher organisms cannot be explained by the diversity of transcription factors alone.[12] We must consider the more complex combinatorial interactions between transcription factors and the *cis*-regulatory elements in higher eukaryotes.

cis-Regulatory Elements

The role of a transcription factor is often determined not by the factor itself, but on which genes (target genes) its cognate binding sites occur.[13] These binding locations (*cis*-regulatory elements) have been traditionally categorized by their function or genomic organization into promoters, enhancers, silencers, and insulators (Figure 1). Promoters are divided into core and proximal promoters depending on their binding partners and distance from transcription start sites (TSS). Core promoters are directly adjacent to a TSS and act as binding sites for general transcriptional factors such as the RNA polymerases II (pol II) and

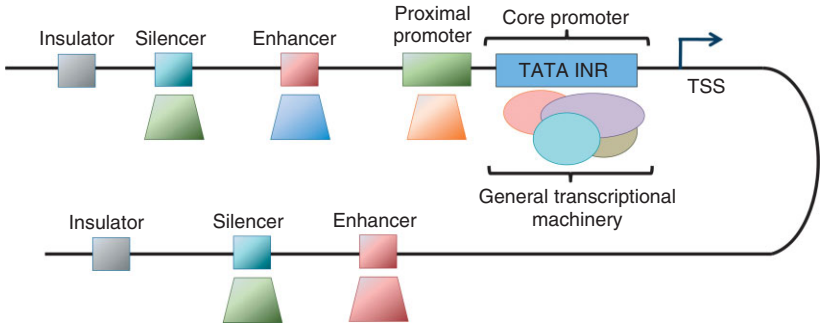


FIGURE 1 | A composite model of *cis*-regulatory elements. Key *cis*-regulatory elements are shown. The core promoters directly upstream of TSS bind the general transcriptional machinery (TATA and INR; refer to the TATA box and the initiator sequence, respectively); proximal promoters bind cognate sequence-specific transcription factors (shown as trapezoids). Distal enhancers and silencer elements also provide binding sites for sequence-specific transcription factors, but have opposite regulatory effects on the transcription activity of regulated genes. In some cases, an enhancer can regulate distant genes, mediated by a DNA-looping mechanism.

© 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

transcription factor II (TFII) families.[14] Unlike the sequence-specific transcription factors, these general transcription factors are generally sequence agnostic. The proximal promoters are located close to the core promoter (within <1 kb) and are bound by sequence-specific transcription factors. The interaction between proximal promoters and transcription factors stimulates transcription by stabilizing the binding of general transcription factors to nearby core promoters.[15] Enhancers also promote transcription; however, they are often located far from the promoters and TSS. They enhance the accessibility of promoter regions by recruiting histone-modifying enzymes that open the chromatin structure.[16] Silencers work similarly to enhancers, but instead suppress transcription. Insulators surround the coding sequence and the *cis*-regulatory elements to protect the gene and regulatory structures from the influence of outside effects. Insulators have also been linked to the organization of chromatin in the nucleus.[17]

It is perhaps surprising that some *cis*-regulatory elements are capable of regulating very remote genes.[18,19] This effect can be understood by considering the spatial separation of features in the nucleus rather than the genomic distance separating them. These two measures may differ because chromatin is not laid out linearly; rather, it is packed in tightly looped structures that may cause distal features in the sequence to be close in the nucleus. This long-range regulatory mechanism involving DNA looping often appears as interchromosomal *trans*-regulating effects.[20] Some recent theories on the structure of chromatin in the nucleus hypothesize that these long-range regulatory effects are not unintentional; instead, they may constitute a higher level of gene regulation.[21,22] There are already sophisticated techniques to investigate regulatory effects by DNA looping[23] and it has been shown that genome-wide mapping tools can be adopted to detect these interactions.[24]

Despite these observations and a number of genome-wide studies suggesting that *cis*-regulatory elements often occur outside the promoter regions,[25] it is generally accepted that functional regulatory elements are primarily located near the TSS of the gene they regulate.[26,27] As a result, the proximal promoter regions have been the traditional target of gene-centric promoter analyses.[28]

DISCOVERY OF MOTIFS AND REGULATORY MODULES

As the majority of transcription factors are sequence-specific,[28] there have been extensive efforts to collect the binding sequences for known transcription factors and to identify potential binding sites in unannotated genome sequence (usually gene promoters). It was initially assumed that an appropriate model for binding patterns could predict whether a specific sequence could bind a given transcription factor. However, as discussed below, sequence specificity does not provide sufficient information to completely determine transcription factor binding sites.

Sequence Motifs

Given a transcription factor, one may be interested in representing its sequence specificity. As the binding domains are generally tolerant to certain variability in the binding sequences, the most simple and straightforward representation is a consensus sequence using IUPAC codes for polymorphic nucleotides. However, this representation clearly lacks important quantitative information: an R encodes either an A or G, but does not indicate the frequency of the two possibilities. To address this issue, matrix representations recording the frequency (position frequency matrix; PFM) and probability (position weight matrix; PWM) for each of the four possible nucleotides for each position in the binding sequence were adopted.[29] These representations can be seen in public databases such as TRANSFAC (http://www.biobase.de/)[30] and JASPAR (http://jaspar.cgb.ki.se).[31] An example of PFM for the c-Myc transcription factor is shown in Figure 2(a). Some groups have proposed more advanced models that include more information per position,[32,33] i.e., by recording dinucleotide (or *n*-nucleotide) frequencies at each position. However, in practice it is generally sufficient to use the standard PWM presentation.[34]

With a known binding motif, we can determine candidate binding positions in a given region of sequence. Given a PWM and a nucleotide sequence of equal length, a binding score can be computed by summing the probability (provided by the PWM) at each position in the queried sequence. Although no functional information can be inferred about the candidate sites, the calculated score agrees well with physical properties such as binding energy.[35] There is an enormous number of algorithms available which try to predict binding sites using the *a priori* knowledge afforded by PWM; a partial list can be found in a recent review.[36]

Unfortunately, binding site prediction based on binding motifs often suffers from high false-positive rates. Wasserman et al. estimated that the number of predicted binding locations for the factor MyoD was on the order of 10^6; however, only 10^3 were estimated to be functional.[36] Although this case is atypical (some reports have shown that motif-based prediction

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

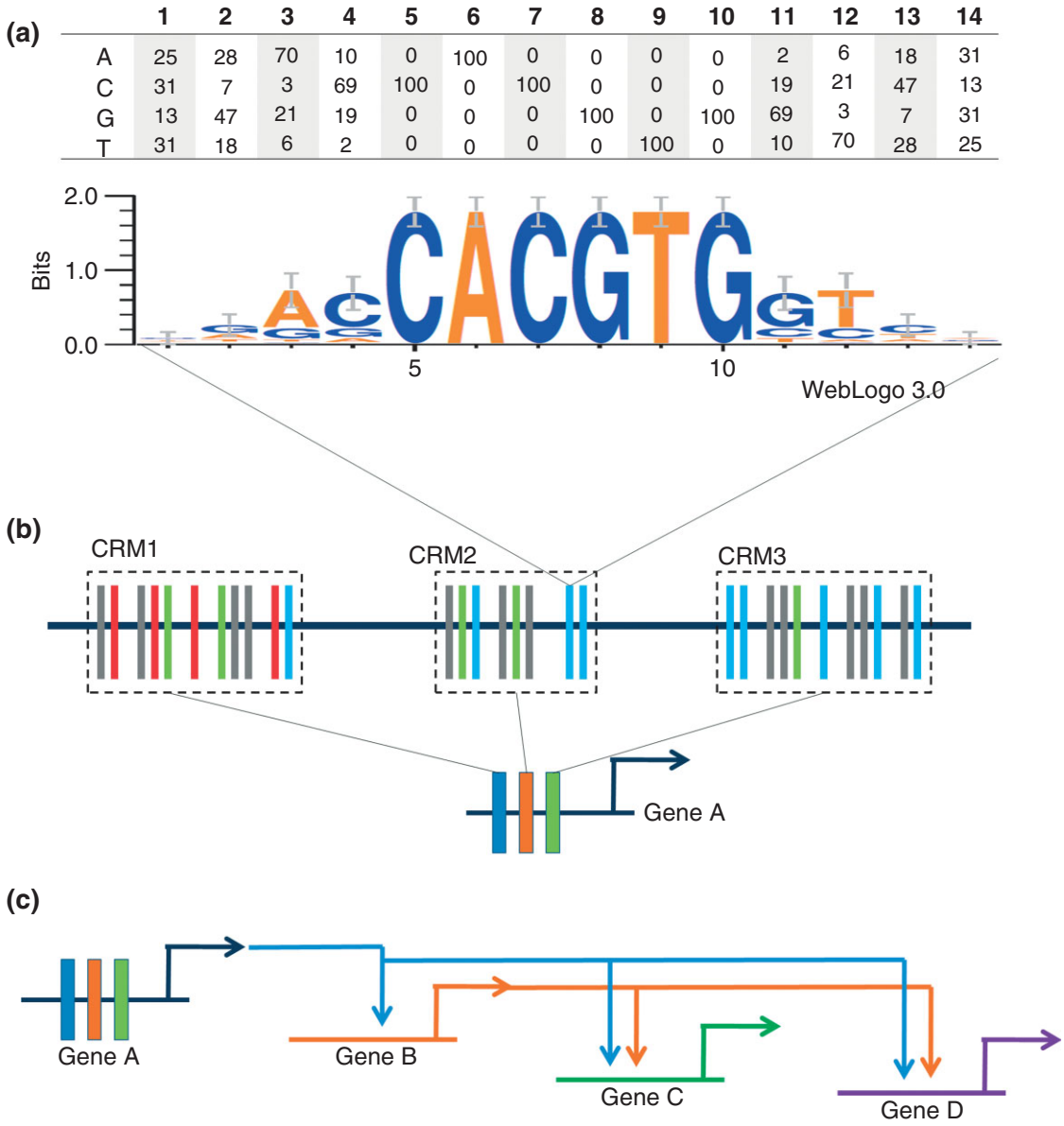


FIGURE 2 | Modeling of individual transcription factor binding sites and *cis*-regulatory modules (CRMs). (a) The position frequency matrix of the c-Myc binding motif (TRANSFAC matrix index M00118, V$MYCMAX_01) is shown as an example. (b) The cluster of individual transcription factor binding sites is modeled as a CRM. CRMs have a different composition and organization of transcription factor binding sites and can be considered as functional units of *cis*-regulation. (c) An example of a regulatory relationship among four genes is shown. Gene A regulates the transcriptional activity of its target gene B and genes A and B together regulate their target genes C and D.

is largely reproducible in *in vitro* tests),[37] it is generally accepted that methods depending only on the knowledge afforded by PWMs will contain a substantial number of false positives. To reduce the false-positive rate, a number of biological insights have been used to integrate extra information into the prediction process. Among these are the evolutionary pressure placed on these important *cis*-regulatory elements; co-expression with genes that have well-documented functions and expression patterns; and the clustering of *cis*-regulatory features into *cis*-regulatory modules (CRMs). We examine these methods below.

Phylogenetic Footprinting

Like coding sequences, functional regulatory elements are under selective pressure during evolution.[38] In some cases, the conservation level of regulatory sequences is even higher than that of the associated coding sequence and surrounding sequences.[39] It therefore seems feasible to filter out regions with

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

low evolutionary conservation from the input of the motif discovery algorithm. This process is referred to as phylogenetic footprinting. It was initially claimed that this filtering method would drastically reduce false-positive rate with a minimal impact on sensitivity.[40,41] Early studies confirmed this claim in comparisons between yeast species[42,43] and between human and rodent.[44] However, unrelated studies later estimated that the conservation level of transcription factor binding sites between human and rodent could be as low as 30–40%, raising suspicion that phylogenetic footprinting should have a substantial impact on sensitivity.[45] This suspicion was confirmed by studies showing that the mutation rate of known regulatory elements was much higher than previous expectations.[46,47] Despite these shortcomings, phylogenetic footprinting has been widely used for the past decade.

Expression Microarray Data

Expression microarrays permit the simultaneous measurement of transcription levels for every known gene in an organism. By collecting measurements of expression levels across a range of cellular conditions and applying clustering techniques, groups of genes with similar expression patterns can be identified.[48] Genes belonging to the same cluster are often called co-expressed. Co-expression was initially used to assign molecular functions to unannotated genes under the assumption that genes with matching expression patterns might share similar molecular functions.[49] This idea was eventually applied to regulatory mechanisms, i.e., genes with similar expression patterns might share transcription factors and functional regulatory binding sites.[50] Coupling this insight with the power of microarrays gave rise to *de novo* motif discovery.

De novo motif discovery algorithms search for short, statistically overrepresented motifs in a set of sequences (generally promoters of co-expressed genes).[51,52] Simulation studies found these algorithms to have a sensitivity of roughly 10%.[53] Despite the low sensitivity, these algorithms have proven useful for studying uncharacterized transcription factors whose PWM is not available.

It is often possible to associate the motifs with dominant functions associated to their gene cluster.[54,55] Segal et al. extended this idea by noting that transcription factors are themselves the product of gene expression; hence, the regulatory relationship with their target genes can be inferred from microarray data.[56] Tavazoie and Beer showed that the expression patterns of individual genes can be predicted by the specific organization of transcription factor binding sites in promoter regions.[57]

Module-Based Approaches to Find Functional Combination

cis-Regulatory elements are often observed in clusters. Current theory explains this propensity for clustering as an organizational feature that facilitates combinatorial interaction between transcription factors. Studies have observed interplay between clustered *cis*-regulatory elements: the regulatory cluster of the *Endo16* gene in sea urchin is a famous example.[58] These clusters of regulatory elements are often viewed as functional units and are generally referred to as CRMs.[59] The general characteristics of a CRM are organism-dependent. For example, CRMs in yeast are typically 300 base pair (bp) in size and contain at least 10 binding sites recognized by at least four different transcription factors[60]; however, a typical CRM in fruitfly is on the order of a kilobase in size.[61] Additional characteristics of the module may be significant as well: Tavazoie has proposed that the exact layout, frequency, and orientation of *cis*-regulatory elements within the module may enact regulatory control.[57,62] Figure 2(b) shows a schematic presentation of CRM as cluster of transcription factor binding sites. Individual genes contain several CRMs, each of which receives and processes upstream regulatory signals and contributes to the final transcription rate of the gene (Figure 2(c)).

Unsurprisingly, functional modules of *cis*-regulatory elements tend to be easier to discover than functional *cis*-regulatory elements. Therefore, CRMs have also proven useful in biological network construction by reducing the number of regulatory elements to be considered. Module-based regulatory networks have been successfully constructed in several model organisms, e.g., fruifly[63] and sea urchin.[64] One classic analysis of the body patterning program in fruitfly found more than 50% of known functional modules and predicted a number of novel modules. Half of the novel modules were tested experimentally and roughly half of these were found to be functional.[63] A number of CRM discovery algorithms are summarized along with their validation rates in a recent publication[65]; however, it is important to note that these rates have been largely derived from tests on simple model organisms. Some critics have claimed that such simple tests do not necessarily imply good performance in more complex organisms.

GENOME-WIDE IDENTIFICATION OF *cis*-REGULATORY ELEMENTS

Chromatin Immunoprecipitation

ChIP (Figure 3) is a method for the identification of DNA fragments that were bound to a protein

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

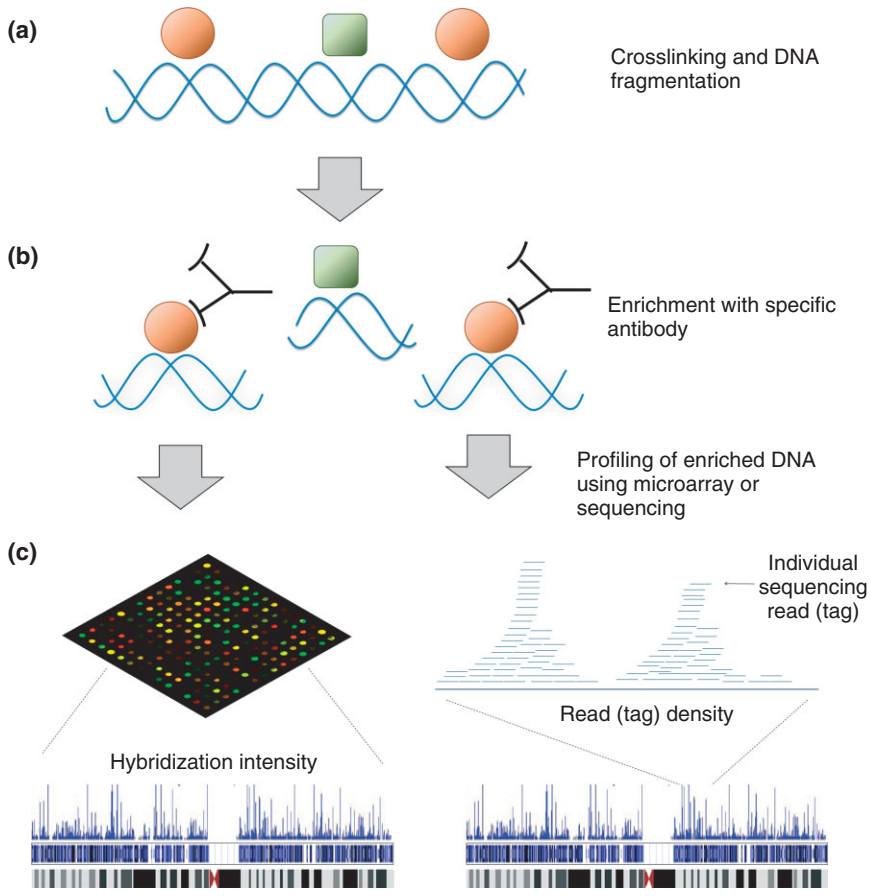


FIGURE 3 | Overview of chromatin immunoprecipitation (ChIP)-chip and ChIP-seq. (a) Reversible cross-linking of DNA and protein is performed by treating the DNA–protein complex with formaldehyde. The cross-linked DNA–protein complex is fragmented by sonication. (b) An antibody specific to the protein of interest is used to enrich the DNA segments bound to the protein. (c) The purified DNA is profiled using a microarray (ChIP-chip; left) or direct sequencing (ChIP-seq; right).

of interest *in vivo*. This technique can be used to profile either a DNA-binding protein or a histone modification, but we will focus on the case of DNA-binding proteins here. First, cell extract is treated with formaldehyde to cross-link the proteins to their binding sites. Then, the cross-linked chromatin is broken up into small fragments (generally in the 200–500 bp range), most often via sonication. The DNA fragments cross-linked to the protein of interest are immunoprecipitated using an antibody specific to that protein, although a better alternative, when possible, is to generate a tagged version of the protein and use the antibody against the tag. After purification, the binding of proteins to DNA is reversed, and the enriched fragments are isolated. Depending on the quality of the antibody, this process commonly yields purified DNA with 5- to 50-fold enrichment for regions of DNA bound by the protein of interest. When combined with a technique to determine which regions of DNA have been selected by immunoprecipitation, this approach allows one to directly identify DNA–protein-binding sites. Compared to motif discovery methods that attempt to infer binding sites indirectly by PWMs and co-expression clusters, ChIP yields the binding sites that are functional in a particular cell type and condition. It is important to note that the specificity of the antibody is critical for generating high-quality data.

ChIP-chip

ChIP-chip combines immunoprecipitation and microarrays to determine the binding sites for a protein of interest. A modern oligonucleotide array can accommodate hundreds of thousands to millions of probes on its surface and these probes can be 'tiled' across specified regions. For small genomes, the whole genome can be tiled at high resolution (e.g., a probe at every 10–100 bp); for large genomes, only selected regions can be covered. A popular version has been the promoter array, which covers the upstream region, e.g., 1–10 kb, of the TSS for every gene and can be useful for examining transcription factor binding.

ChIP-chip was the first method for genome-wide mapping of DNA–protein interaction *in vivo*, and provided an opportunity for the first time to examine the regulator–gene interactions in a global regulatory network. It was applied in multiple studies on a wide range of organisms and cell types (from yeast[66] to human embryonic stem cells[67]) revealing

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

network architecture for the first time. As the microarray technology improved, the quality of data has also improved. Nearly a decade ago, when ChIP-chip was first introduced with cDNA microarrays, two independent studies in yeast focusing on the same transcription factor complexes showed a 30% agreement between predicted binding sites.[68,69] In a recent comparison of microarray platforms, Johnson et al.[70] created mock ChIP DNA by enriching a set of known DNA sequences at various concentrations. This mock ChIP DNA was then analyzed by three widely used microarray platforms to determine the enriched sites. Despite the differences in the platforms, performance was fairly uniform: each platform discovered approximately 50% of the enriched sites with a false discovery rate of 5%.

ChIP-seq

Massively parallel sequencing, often referred to as 'next-generation sequencing' (NGS), is a recent innovation that takes advantage of sequencing-by-synthesis (Illumina, San Diego, CA and Helicos, Cambridge, MA), pyrosequencing (454 Life Sciences, Branford, CT), and sequencing-by-ligation (Applied Biosystems, Foster City, CA) to generate hundreds of millions of short reads in a single run (several hundred bp for 454 and <100 bp for the rest).[71,72] Although NGS has been applied to numerous areas, ChIP-seq has been one of its most successful applications. In ChIP-seq, conventional ChIP is used to enrich the DNA fragments corresponding to sites of DNA–protein interactions, but it is followed by NGS to directly sequence the fragments. The sequenced reads are then aligned to the genome and one of the several peak-finding algorithms can be used to determine the likely binding locations.[73] The process is illustrated in Figure 3. One of the main disadvantages of NGS in general is that the reads are much shorter than in conventional Sanger sequencing (800–1000 bp). This means that the reads are less likely to map uniquely to the genome and there is a smaller percentage of the genome to which a read can be aligned. Fortunately, unlike *de novo* genome assembly or some other applications, the ChIP fragments to be sequenced are small (e.g., 200–500 bp) and ChIP-seq is affected only slightly by the shorter length of the current NGS reads.

In mapping transcription factors for construction of a regulatory network, there are two primary advantages to ChIP-seq. First, there is an increased resolution in the binding sites detected. In ChIP-chip, the resolution of the binding profile is dictated by the spacing between tiled probes on the array. Although it depends on the design of a particular array, the resolution is typically >100 bp. In ChIP-seq, the sequenced reads can be aligned to nearly any location in the reference genome, and thus the binding profile can be single-base-pair resolution in theory. In practice, the distribution of the ChIP DNA fragment lengths becomes the limiting factor for spatial resolution, assuming enough reads have been sequenced to distinguish the peaks clearly. The higher resolution of ChIP-seq can be important for accurate detection of binding motifs. In the study by Johnson et al.,[74] the authors aimed to identify the neuron-restrictive silencer factor (NRSF)-binding sequence motif that consisted of two half-sites separated by a gap of variable size. It was only with the increased resolution of ChIP-seq that this motif could be derived.

The second advantage is that the experiment is cost-effective for large genomes. For tiled microarrays, binding of a transcription factor can be interrogated only for the pre-defined regions of the genome. Covering the entire human genome on microarrays at ~100 bp resolution requires roughly 30 arrays,[70] which is prohibitively expensive. For ChIP-seq, the cost of the experiment does not depend on the genome size, although more reads would be necessary for large genomes because they are likely to have more binding sites per factor in general. A recent paper describes a relationship between the number of reads (sequencing depth) and peak properties (resolution, enrichment, and number) in ChIP-seq studies.[75]

Some NGS technologies also require less DNA than microarrays. A recent paper demonstrated that ChIP-seq can be done with as little as 50 pg on a single-molecule sequencing platform.[76] This is in contrast to several micrograms needed for microarrays, typically from millions of cells. In many studies, the number of cells that can be obtained for ChIP is limited; thus, the ability to perform experiment with a smaller amount of starting material can be critical.

Although ChIP-seq is a superior alternative to ChIP-chip in general, it does require sufficient depth of sequencing to identify the binding sites comprehensively. When the number of tags is too small, weaker peaks will be missed and the false-negative rate may be high. Exactly how many tags are necessary is a difficult question and the answer depends on the shape and the number of enriched regions. But some estimates can be derived by simulation and interpolation based on the tags already sequenced.[77] The NGS technology also has some artifacts of its own. Its base-calling error rate is higher than that of traditional Sanger sequencing and this must be considered in the alignment step. There are other biases that may occur during amplification, library preparation, or the actual sequencing in the machine. In particular, there is GC bias, i.e.,

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reads produced by NGS platforms often have a higher average GC content than expected,[78,79] meaning that fragments with higher GC content are preferentially selected for sequencing. Many of these biases can be addressed by generating a control sample using input DNA, which goes through exactly the same steps as the ChIP samples but without immunoprecipitation.

A NETWORK PERSPECTIVE ON TRANSCRIPTIONAL REGULATION

There are several types of biological networks, depending on elements and interactions that serve as nodes and edges, respectively. Well-known examples include metabolic, protein–protein interaction, and transcriptional regulatory networks. Although reconstruction of these networks has been hampered by a lack of high-quality and comprehensive data, recent high-throughput methods have improved significantly and have facilitated a better understanding of the data in a network- or systems-level context. Accompanying the advances in experimental techniques have been developments in the analysis of complex networks to describe the organizing principles of biological and social networks. In this section, we will address some network-level properties of transcriptional regulatory networks and discuss the insights that can be gleaned from this type of analysis.

Organization of Transcriptional Regulatory Networks

At a global scale, networks can be categorized into three classes according to their topology: random networks, scale-free networks, and hierarchical networks.[80] The distribution of node degree (connectivity) has been used to distinguish these networks. In scale-free networks, the distribution of node degree follows a power-law distribution, i.e., the number of nodes with k edges is approximately $k^{-\gamma}$, where γ is a constant that depends on the network.[81] Many biological (e.g., metabolic and interaction networks) as well as non-biological networks (e.g., social networks and the World Wide Web) appear to follow this scale-free topology.[80,82–84] In transcriptional regulatory networks, connectivity is directional and can be divided into outgoing (the number of target genes regulated by each transcription factor) and incoming connectivity (the number of transcription factors regulating a target gene). Guelzim et al.[85] observed that outgoing connectivity in yeast transcriptional regulatory network was shown to fit a power-law distribution, suggesting that a small set of transcription factors (regulatory hubs) has an exceptionally high number of connections, whereas the vast majority are minimally connected.[66,85] Interestingly, the incoming connectivity follows an exponential decay after a peak at a mean, indicating that most of the target genes have a small number of transcriptional regulators (e.g., 93% of genes are regulated by one to four transcription factors). The scale-free topology of outgoing connectivity (i.e., presence of regulatory hubs) can provide robustness to a system with increased resilience to random perturbations, but it may be vulnerable to targeted attack at critical regulatory hubs.[86] Such critical failures have been observed in yeast, where the removal of regulatory hubs tended to be more lethal.[87] Many scale-free networks also show an 'ultra-small world' property in which only a small number of steps are required to join two nodes.[88] Taken together, the topology of transcriptional regulatory networks provides robustness and ensures efficient signal transduction within the network.

At a local level, a network can be decomposed into subunits called motifs or modules. Network motifs are the basic building blocks of a network[89] and they represent small-scale interaction patterns that describe how nodes are connected with their neighbors. In transcriptional regulatory networks, certain motifs between genes have been found to appear at a higher frequency than can be found in randomized networks of similar size.[90] Figure 4(a) illustrates the example of such motifs: autoregulation, feed-forward loops, and single- or multi-input motifs.[66] Network motifs can provide mechanistic insights regarding how signals are propagated and integrated. Feed-forward loops, for instance, are robust to random fluctuation because the target gene expression is regulated by two transcription factors; single- and multi-input motifs have been associated with co-expression of multiple target genes and the integration of diverse signals via multiple transcription factors, respectively.

Some network motifs such as feed-forward loops have been conserved in evolution, supporting the functional relevance of network motifs.[90] From an evolutionary perspective, it was initially assumed that a gene duplication event is a major driving force in the creation of network motifs for a transcriptional regulatory network, i.e., the duplicated genes will inherit the interactions of their ancestors.[91] This might be the case for less diverged genes with recent duplication histories,[92] but detailed examination of known network motifs in transcriptional regulatory networks in yeast and *Escherichia coli* has shown that the extent to which common ancestry is shared is not higher than expected by chance. This suggests that motifs in transcriptional regulatory networks follow their own, convergent evolutionary program.[93] It is also of note

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

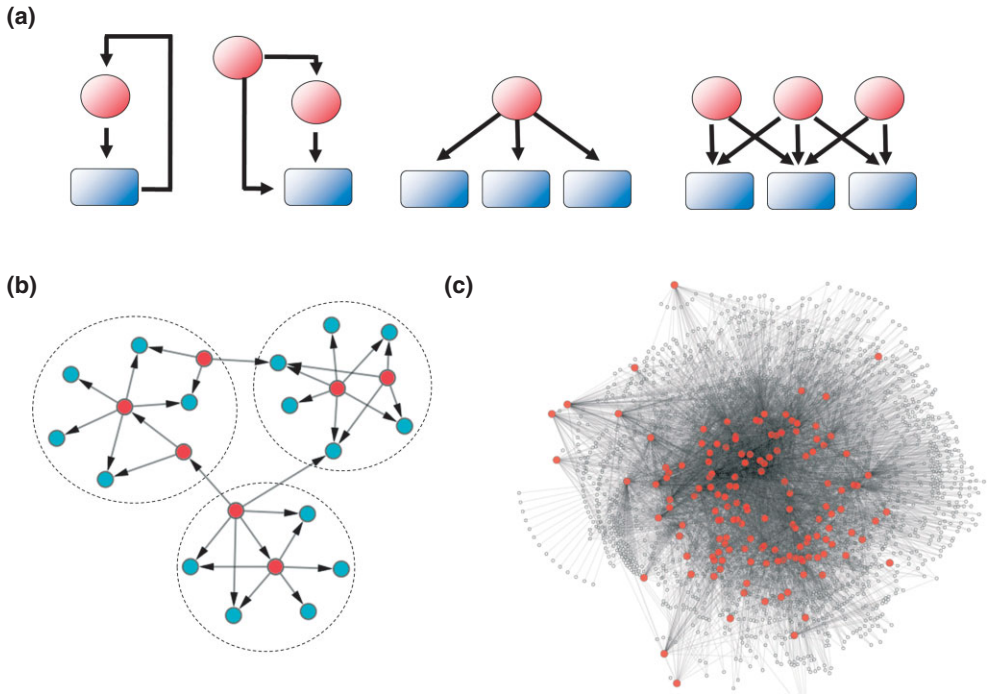


FIGURE 4 | Examples of network motifs, modules and transcriptional regulatory network. (a) Well-known network motifs in transcriptional regulatory networks are shown. These are autoregulation, a feed-forward loop, and single- and multiple-input motifs[66] (left to right). The red circles and blue rectangles represent the transcriptional factors (regulators) and their target genes, respectively. (b) An example of modules in a network is shown. The nodes within the modules are highly interconnected with each other. A clustering coefficient can be calculated to measure the extent of connectivity or cliqueness of nodes in a module. (c) A yeast transcription regulatory network was constructed here using 17,873 known relationships between transcription factors and their target genes. The genes were classified into transcription regulators (red circle) and regulated genes (black circle) according to the annotations in the original manuscript.[95] The network is visualized using the Cytoscape software with the organic layout option.[98]

that the network characteristics such as motifs and topology can dynamically change according to cellular conditions. For example, Luscombe et al.[94] reported that such re-wiring of yeast transcriptional regulatory network is highly optimized for required cellular functions, both for fast signal propagation in exogenous functions (e.g., stress response) and for temporal progression with highly interconnected transcription factors in endogenous functions (e.g., cell cycle).

Highly interconnected nodes or motifs can comprise higher-order network modules (Figure 4(b)). Figure 4(c) shows a yeast transcriptional regulatory network with 17,873 known transcriptional regulatory relationships.[95] The extent of node connectivity within a module or within an entire network can be measured by the average clustering coefficient[80] and it was shown that yeast transcriptional regulatory networks have fivefold higher clustering coefficients than expected by chance.[85] Thus, the transcriptional regulatory network appears to be modular, consistent with the assumption that modularity is a universal design principle in biological systems.[82,96] In addition, the separation by modules in network topology raised the possibility that these modules represent functionally independent units exerting particular biological functions. This notion originated from the co-expressed gene modules obtained by conventional clustering methods, but a number of algorithms have been proposed to define modules directly from the network using hierarchical or graph clustering.[97]

Integration with Other Biological Networks

A metabolic network is a well-established type of biological network. Compared to a transcriptional regulatory network, a metabolic network can be more easily reconstructed for an organism (whose genome sequence is available), given that biochemical relationships between key metabolic genes have been established.[99] This is due to the higher level of sequence conservation of key metabolic enzymes (nodes) as well as the fixed relationships between them (edges) in well-characterized metabolic pathways. The prediction of nodes (*cis*-regulatory elements) and edges (DNA–protein interactions) in transcriptional regulatory networks using evolutionary constraints,

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

on the other hand, is limited. Integration of these two types of networks has proven successful in prokaryotes. For example, Covert et al.[100,101] constructed a combined metabolic–regulatory model by incorporating the known relationships between transcription factors and metabolic enzymes onto the metabolic network of *E. coli*. Their model was able to accurately predict the altered metabolic behaviors under the influence of different transcription factors or external signals. This type of integration is challenging, especially due to the differing reliability and interpretation of network edges. For proper integration of networks, discrete Boolean networks or probabilistic Bayesian networks can be used to deal with a combination of different types of interactions (transcriptional regulation or metabolic pathway) occurring on the same gene.[102]

One of the assumptions underlying the clustering of binding sites in CRMs is that multiple transcription factors interact in complex, combinatorial patterns. Given that methods for detection of protein–DNA interactions only provide indirect evidence for combinatorial effects, integration of data from experimentally validated protein–protein interactions (PPI) is likely to give additional information on the combinatorial action among transcription factors. In addition, PPI databases may provide clues for long-range interactions in transcriptional control that are mediated by PPI, e.g., interactions between transcription factors and cofactors.[103] The main PPI databases include the Human Protein Reference Database (http://www.hprd.org/) and the Munich Information Center for Protein Sequences (http://mips.gsf.de).[104,105] Although these databases are biased toward well-known genes, it is useful to examine the network structure based on these data. For example, the proximity of genes in a PPI network represents a putative functional relationship, i.e., neighboring genes in the network are likely to share the same biological function.[106] This insight has been used successfully to identify novel functions for unannotated genes as well as to discover novel disease genes. Examples include *GIT1*, which was found to be associated with Huntington's disease by being in the same PPI subnetwork with the known disease gene *HTT*,[107] and *Puratrophin-1*, which was identified as one of the novel ataxia-related genes by the same approach.[108] This idea also can be used to infer functional relationships of transcription factors in a transcriptional regulatory network. For example, in the network describing the *C. elegans* digestive tract, Deplancke et al. observed that two candidates (DIE-1 and ZTF-2) that share many target genes indeed have similar binding motifs and protein structures.[109,110]

In a co-expression network, the correlation between a pair of genes can be measured across diverse cellular conditions, and potential relationship (edges) between genes (nodes) is inferred for those with a high correlation.[111] Although simple correlation does not guarantee a functional relationship between genes, studies have reported evolutionary conservation of such relationships and that they can be used to infer gene function.[112,113] These co-expression relationships can be integrated into the reconstruction of transcriptional regulatory networks. For this purpose, there are a number of reverse-engineering network-inference algorithms such as ARACNE,[114] MRNET,[115] and CLR.[116] These algorithms were recently benchmarked with different correlation estimators.[117,118] One of the popular algorithms, ARACNE (algorithm for the reconstruction of accurate cellular networks), has been used to infer transcriptional regulatory networks from large-scale expression data sets, e.g., the c-Myc-related subnetwork in human B-cells[119] as well as a network of STAT3 and CEBPβ transcription factors from cancer transcriptome.[120] These successes highlight the potential use of large-scale expression data to evaluate and supplement the construction of transcriptional regulatory networks.

CONCLUSION

In the past few decades, we have learned a great deal about global transcriptional regulatory networks in small organisms (e.g., prokaryotes and yeast). There are also detailed examples of specific transcriptional regulatory networks in higher model organisms[60,121] that illustrate how *cis*-regulatory elements integrate inputs from multiple transcription factors and control the gene output. However, how such controls are constructed in more complex organisms such as humans remains largely elusive. Previous motif- or CRM-based approaches, e.g., predicting tissue-specific genes using organization of CRMs in the promoter regions,[65,122] in a more complex genome have only resulted in limited insights.

Fortunately, recent developments in high-throughput sequencing technologies have added a new set of tools for studying regulatory relationships. In particular, ChIP-seq now allows one to accurately identify genome-wide targets of a transcription factor in higher eukaryotes with reasonable cost, and RNAi knockdown and other large-scale screening methods allow faster identification of key components. Generation of sufficient amounts of data in diverse conditions, tissue types, and time points remains a significant challenge, but these tools will facilitate

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

a comprehensive identification of the transcriptional factors and their interactions with *cis*-regulatory elements. With appropriate computational methods, these data will enable a more faithful construction of transcriptional regulatory networks and more accurate predictions of cellular responses to diverse signals.

REFERENCES

1. Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 2000, 34: 77–137.
2. Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000, 14:2551–2569.
3. Wray GA. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 2007, 8: 206–216.
4. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006, 7:29–59.
5. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001, 409:853–855.
6. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009, 10:252–263.
7. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* 1989, 245:371–378.
8. Kummerfeld SK, Teichmann SA. DBD: a transcription factor prediction database. *Nucleic Acids Res* 2006, 34:D74–D81.
9. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004, 431:99–104.
10. Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout AJ. A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol* 2005, 6:R110.
11. van Nimwegen E. Scaling laws in the functional content of genomes. *Trends Genet* 2003, 19:479–484.
12. Levine M, Tjian R. Transcription regulation and animal diversity. *Nature* 2003, 424:147–151.
13. Madan BM, Teichmann SA. Functional determinants of transcription factors in Escherichia coli: protein families and binding sites. *Trends Genet* 2003, 19:75–79.
14. Thomas MC, Chiang CM. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* 2006, 41:105–178.
15. Gill G. Regulation of the initiation of eukaryotic transcription. *Essays Biochem* 2001, 37:33–43.
16. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 2009, 10:161–172.
17. Gerasimova TI, Byrd K, Corces VG. A chromatin insulator determines the nuclear localization of DNA. *Mol Cell* 2000, 6:1025–1035.
18. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. *Science* 2003, 302:413.
19. Liu Z, Garrard WT. Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3′ boundary sequence spanning 46 kilobases. *Mol Cell Biol* 2005, 25:3220–3231.
20. Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA. Interchromosomal associations between alternatively expressed loci. *Nature* 2005, 435:637–645.
21. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2001, 2:292–301.
22. Lamond AI, Earnshaw WC. Structure and function in the nucleus. *Science* 1998, 280:547–553.
23. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science* 2002, 295: 1306–1311.
24. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009, 326:289–293.
25. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* 2005, 434:338–345.
26. Bieda M, Xu X, Singer MA, Green R, Farnham PJ. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* 2006, 16:595–605.
27. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 2006, 16:656–668.
28. Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 2004, 116:247–257.

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16:16–23.
30. Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996, 24:238–241.
31. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2010, 38:D105–D110.
32. Hannenhalli S, Wang LS. Enhanced position weight matrices using mixture models. *Bioinformatics* 2005, 21(suppl 1):i204–i212.
33. Georgi B, Schliep A. Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics* 2006, 22:e166–e173.
34. Benos PV, Bulyk ML, Stormo GD. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* 2002, 30:4442–4451.
35. King OD, Roth FP. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res* 2003, 31:e116.
36. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 2004, 5:276–287.
37. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J Mol Biol* 1997, 266:231–245.
38. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* 1988, 203:439–455.
39. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol* 2003, 3:19.
40. Hardison RC. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* 2000, 16:369–372.
41. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 2005, 3:e7.
42. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 2003, 301:71–76.
43. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 2003, 423:241–254.
44. Levy S, Hannenhalli S. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* 2002, 13:510–514.
45. Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 2002, 19:1114–1121.
46. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* 2008, 6:e27.
47. King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, Chiaromonte F, Miller W, Hardison RC. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* 2007, 17:775–786.
48. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998, 95:14863–14868.
49. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al. Functional discovery via a compendium of expression profiles. *Cell* 2000, 102:109–126.
50. Yu H, Luscombe NM, Qian J, Gerstein M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* 2003, 19:422–427.
51. Bailey TL, Elkan C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 1995, 3:21–29.
52. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000, 296:1205–1214.
53. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005, 23:137–144.
54. Dieterich C, Rahmann S, Vingron M. Functional inference from non-random distributions of conserved predicted transcription factor binding sites. *Bioinformatics* 2004, 20(suppl 1):i109–i115.
55. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res* 2003, 13:773–780.
56. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003, 34:166–176.

57. Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell* 2004, 117:185–198.
58. Yuh CH, Davidson EH. Modular cis-regulatory organization of Endo16, a gut-specific gene of the sea urchin embryo. *Development* 1996, 122:1069–1082.
59. Jeziorska DM, Jordan KW, Vance KW. A systems biology approach to understanding cis-regulatory module function. *Semin Cell Dev Biol* 2009, 20:856–862.
60. Levine M, Davidson EH. Gene regulatory networks for development. *Proc Natl Acad Sci U S A* 2005, 102:4936–4942.
61. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu XJ, White KP, Bussemaker HJ, et al. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proc Natl Acad Sci U S A* 2006, 103:12027–12032.
62. Erives A, Levine M. Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci U S A* 2004, 101:3851–3856.
63. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* 2004, 2:E271.
64. Ben-Tabou de-Leon S, Davidson EH. Deciphering the underlying mechanism of specification and differentiation: the sea urchin gene regulatory network. *Sci STKE* 2006, 2006:e47.
65. van Loo P, Marynen P. Computational methods for the detection of cis-regulatory modules. *Brief Bioinform* 2009, 10:509–524.
66. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002, 298:799–804.
67. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 2005, 122:947–956.
68. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* 2001, 106:697–708.
69. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001, 409:533–538.
70. Johnson DS, Li W, Gordon DB, Bhattacharjee A, Curry B, Ghosh J, Brizuela L, Carroll JS, Brown M, Flicek P, et al. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res* 2008, 18:393–403.
71. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, et al. The challenges of sequencing by synthesis. *Nat Biotechnol* 2009, 27:1013–1023.
72. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008, 26:1135–1145.
73. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009, 10:669–680.
74. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007, 316:1497–1502.
75. Barski A, Zhao K. Genomic location analysis by ChIP-Seq. *J Cell Biochem* 2009, 107:11–18.
76. Goren A, Ozsolak F, Shoresh N, Ku M, Adli M, Hart C, Gymrek M, Zuk O, Regev A, Milos PM, et al. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods* 2010, 7:47–49.
77. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008, 26:1351–1359.
78. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008, 36:e105.
79. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 2008, 5:1005–1010.
80. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004, 5:101–113.
81. Barabasi AL. Scale-free networks: a decade and beyond. *Science* 2009, 325:412–413.
82. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002, 297:1551–1555.
83. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature* 2001, 411:41–42.
84. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000, 407:651–654.
85. Guelzim N, Bottani S, Bourgine P, Kepes F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 2002, 31:60–63.
86. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature* 2000, 406:378–382.

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

87. Yu H, Greenbaum D, Xin LH, Zhu X, Gerstein M. Genomic analysis of essentiality within protein networks. *Trends Genet* 2004, 20:227–231.

88. Cohen R, Havlin S. Scale-free networks are ultrasmall. *Phys Rev Lett* 2003, 90:058701.

89. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science* 2002, 298:824–827.

90. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 2002, 31:64–68.

91. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. *Nat Genet* 2004, 36:492–496.

92. Papp B, Pal C, Hurst LD. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet* 2003, 19:417–422.

93. Conant GC, Wagner A. Convergent evolution of gene circuits. *Nat Genet* 2003, 34:264–266.

94. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 2004, 431:308–312.

95. Balaji S, Iyer LM, Aravind L, Babu MM. Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *J Mol Biol* 2006, 360:204–212.

96. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al. A protein interaction map of Drosophila melanogaster. *Science* 2003, 302:1727–1736.

97. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007, 3:88.

98. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13:2498–2504.

99. Herrgard MJ, Covert MW, Palsson BO. Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* 2004, 15:70–77.

100. Barrett CL, Herring CD, Reed JL, Palsson BO. The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. *Proc Natl Acad Sci U S A* 2005, 102:19103–19108.

101. Covert MW, Palsson BO. Transcriptional regulation in constraints-based metabolic models of Escherichia coli. *J Biol Chem* 2002, 277:28058–28064.

102. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002, 9:67–103.

103. Farnham PJ. Insights from genomic profiling of transcription factors. *Nat Rev Genet* 2009, 10:605–616.

104. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2002, 30:31–34.

105. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human Protein Reference Database - 2009 update. *Nucleic Acids Res* 2009, 37:D767–D772.

106. Ideker T, Sharan R. Protein networks in disease. *Genome Res* 2008, 18:644–652.

107. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell* 2004, 15:853–865.

108. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, et al. A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 2006, 125:801–814.

109. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, et al. A gene-centered C. elegans protein–DNA interaction network. *Cell* 2006, 125:1193–1205.

110. Walhout AJ. Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping. *Genome Res* 2006, 16:1445–1454.

111. Carter SL, Brechbuhler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 2004, 20:2242–2250.

112. van Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. *Trends Genet* 2003, 19:238–242.

113. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003, 302:249–255.

114. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006, 7(suppl 1):S7.

115. Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol* 2007, 2007:79879.

116. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 2007, 5:e8.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

117. Olsen C, Meyer PE, Bontempi G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP J Bioinform Syst Biol* 2009, 2009:308959.

118. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* 2008, 9:461.

119. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet* 2005, 37:382–390.

120. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 2010, 463:318–325.

121. Bonn S, Furlong EE. cis-Regulatory networks during development: a view of Drosophila. *Curr Opin Genet Dev* 2008, 18:513–520.

122. Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 1998, 278:167–181.

 © 2010 John Wiley & Sons, Inc.

1939005x, 2011, 1, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/wsbm.105 by Harvard University, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License