Published in final edited form as:
Nat Rev Genet. ; 13(7): 469–483. doi:10.1038/nrg3242.

Genomic approaches to finding *cis*-regulatory modules in animals

Ross C. Hardison
Department of Biochemistry and Molecular Biology, Center for Comparative Genomics and Bioinformatics, 304 Wartik Laboratory, The Pennsylvania State University, University Park, PA 16802, USA rch8@psu.edu

James Taylor
Departments of Biology and Mathematics and Computer Science, Emory University, O. Wayne Rollins Research Center, 1510 Clifton Road NE, Atlanta, GA 30322, USA james.taylor@emory.edu

Abstract

Differential gene expression is the fundamental mechanism underlying animal development and cell differentiation. However, it is a challenge to identify comprehensively and accurately the DNA sequences required to regulate gene expression, called *cis*-regulatory modules (CRMs). Three major features (singly or in combination) are used to predict CRMs: clusters of transcription-factor binding-site motifs, noncoding DNA under evolutionary constraint, and biochemical marks associated with CRMs, such as histone modifications and protein occupancy. The validation rates for predictions indicate that identifying diagnostic biochemical marks is the most reliable method, and understanding is enhanced by analysis of motifs and conservation patterns within those predicted CRMs.

The development of animals from zygotes to adults and the differentiation of cells into distinct tissues and organs requires the expression of a specific set of genes at each developmental stage and in each cell type[1]. The features distinguishing humans from apes have long been attributed to differences in gene expression[2], and aberrant gene expression lies at the heart of multiple diseases. Thus, identifying the DNA sequences required for regulating gene expression, called *cis*-regulatory modules (CRMs), can both expand our understanding of biology and have applications in several fields including evolution and medicine. For example, most of the genetic variants significantly associated with susceptibility to disease do not lie in protein-coding regions[3], and we surmise that many affect the regulation of gene expression.

Three major approaches have emerged for predicting CRMs. The first is to search genomic DNA for clusters of short motifs that are needed for the specific binding of transcription factors (TFs). Although CRMs should contain multiple such motifs, this approach to identifying CRMs has had limited success. A second approach for identifying CRMs involves comparing homologous, noncoding DNA sequences between related species. These methods can reveal important subsets of conserved CRMs that are under **purifying selection**, such as developmental enhancers, but they miss lineage-specific ones. More recently, high-throughput, direct assays for DNA sequences that have epigenetic features characteristic of regulatory regions provide a third approach that has potentially high predictive power for identifying CRMs. This method, which involves mapping the locations of TF-binding and histone modifications in a wide range of tissues and developmental stages, yields an unbiased genomic view of potential gene-regulatory regions that is not restricted to conserved regions or those with known regulatory motifs.

We briefly review the major types of CRMs being studied in animals and then review the strengths and weaknesses of the three approaches to CRM prediction, assessing the success rates of each. We suggest ways to use the three approaches in combination to improve predictions, and discuss important questions for future research. Improvements in CRM prediction and classification are already leading to advances in understanding how genetic variants affect susceptibility to disease[4–7].

Our emphasis in this review is to assess the efficacy of these methods and suggest ways in which they can be improved. Readers are referred to other recent reviews for more details on the biochemical features of chromatin around CRMs[8–13], prediction methods that are based on conservation and motifs[14,15], and earlier comparisons of the different approaches[16,17].

Classes of *cis*-Regulatory Modules

Regulation of gene expression involves an interaction between TFs and CRMs, and it is important to be clear about how one refers to the DNA sequences that TFs can bind (Box 1). In this review we emphasize the **TF binding sites (TFBS)** that are occupied in living cells. The emphasis on *in vivo* occupancy is crucial. Biochemical assays in solution, such as electrophoretic mobility shift assays, *in vitro* footprints, and capture of TF-bound sequences, can define the sequence required for recognition of DNA by TFs; algorithms that assess DNA sequence similarity to a **TFBS motif** will therefore be able to detect millions of motif instances in a mammalian genome[18,19]. While any motif instance could potentially be bound *in vivo*, only about one in 500 actually are bound in organisms with large genomes[18]. As a specific example, the mouse genome contains about 8 million instances of a match to the GATA-binding factor 1 (GATA1) binding site motif, but only about 15,000 DNA segments (some with multiple motif instances) are bound by this transcription factor in erythroid cells[18,20].

CRMs in animal genomes are usually placed into one of three categories, defined by their role in gene expression. The ability of the three prediction methods to detect a CRM depends on the properties of each particular CRM class. This review covers work in both flies (*Drosophila melanogaster*) and mammals because a large number of studies have been done in these species and the fundamental mechanisms of regulation are similar in insects and mammals. However, some genomic features and proteins are present only in one clade, and the smaller size of the fly genome coupled with the lower proportion of noncoding DNA may contribute to a greater success of CRM predictions in this species.

Promoters

A **promoter** directs RNA polymerase to initiate at the transcription start site, TSS[21]. In promoters for RNA polymerase II, general transcription factors bind to a core promoter of about 100 bp around the TSS and facilitate binding of the polymerase complex[8] (**Box 1**). Some core promoters contain well-known motifs such as a TATA box and have a discrete start site for transcription; however, most promoters in mammalian genomes are GC- and CpG-rich regions that lack TATA boxes and tend to support initiation of transcription at a broad range of positions within a roughly 100 bp interval[22]. The heterogeneity in sequence composition and genomic structure of promoters has complicated the accurate prediction of this CRM class based on single sequences. Furthermore, CpG islands are not present in *Drosophila melanogaster*. However, promoters do reside in chromatin with distinctive modifications (**Box 1**), and they can be identified by mapping the start sites for transcription (see below). The functional and mechanistic implications of the differences in promoter classes along with distinct chromatin modifications was recently reviewed[13].

Enhancers and silencers

Enhancers[23–25] and **silencers**[26] are defined operationally by their positive or negative effects, respectively, on a reporter gene after transfer into a transgenic animal or transfected cells in culture. They can act independently of position and orientation in gene transfer assays (**Box 4**). However, depending on the *trans*-environment in a cell, a given DNA segment can switch between enhancing and silencing, presumably reflecting the recruitment of co-activators and co-repressors, respectively[27,28]. Hence the successful prediction of enhancers will probably identify some silencers as well. Currently, few silencers are well-characterized, and they will not be covered further in this review.

Enhancers can be located close to their target promoter[29], but many of them are located a long distance away; an enhancer for the mouse *Shh* gene is 1 Mb away from the *Shh* promoter[30]. An enhancer contains multiple TFBSs (Box 1), and this multiplicity is a requirement for enhancement[31,32]. Genes can have multiple, distinct enhancers that drive expression in specific tissues depending on the particular TFBS motifs and the TFs that bind to them[1,33–35].

The variability in the distance of enhancers from a TSS and the diversity in their composition make prediction of enhancers particularly challenging. The set of mammalian TFs, estimated to be at least 1,000 in number, bind to hundreds of TFBS motifs, but these motifs are short and the vast majority of motif instances are not bound by a TF. As will be developed in a later section, these sequence features are not sufficient for consistently accurate predictions of enhancers. However, including signatures of purifying selection and especially direct evidence for distinctive epigenetic features (**Box 1**) improves the prediction accuracy.

Insulators

Insulators are CRMs that restrict the effect of long-range regulatory modules, such as enhancers, so that they act on the appropriate promoter target[36,37]. One way to do this is via an **enhancer-blocking** activity. When located between an enhancer and a target promoter, such an insulator can block the activity of the enhancer and thereby reduce gene expression[38]. CCCTC-binding factor (CTCF) is a protein required for the enhancer-blocking activity of mammalian insulators[39] (**Box 1**), whereas *Drosophila* species have at least four additional proteins sufficient for enhancer blocking activity, some of which can be identified in other insects[40]. Insulators that serve as **barriers** can prevent **position effects** when they surround a stably integrated reporter gene[41], presumably by blocking the spread of repressive heterochromatin from the site of integration into the reporter gene. This is a separate activity from enhancer blocking, and it requires different proteins such as upstream stimulatory factor (USF), which in turn recruits histone modifying enzymes[42]. The enhancer blocking and barrier activities can occur together in some insulators or separately in others.

As for enhancers, an insulator can be located almost anywhere relative to a gene, and thus location offers no predictive power. Known insulators are located in chromatin with a histone modification profile similar to that of enhancers, but the requirement for CTCF distinguishes enhancer-blocking insulators from enhancers (**Box 1**). A major complication is that CTCF has many additional functions in addition to insulation[43]. Thus, finding CTCF-bound DNA segments should identify most instances of this type of insulator[44], but many of the CTCF-bound segments will not necessarily be insulators. The challenge is to identify those other functions.

Single-genome bioinformatic approaches

The observation that clusters of TFBS motifs are necessary for TF binding to CRMs motivated initial motif-based approaches for predicting enhancers and promoters. The advantage of these approaches is that predictions can be made using only genomic DNA sequence and models of the TFBS motifs for the TFs involved in the process under study (**Box 2**). However, clusters of TFBS motifs occur frequently in large genomes, and alone they are not sufficient for TF binding (e.g. epigenetic marks are required). Thus, genome-wide CRM predictions based on TF motifs typically make many false positive predictions, and consequently have low validation rates. When the search space can be reduced, e.g. by interrogating species with smaller genomes, restricting to relevant genes or using general epigenetic marks, TFBS motif approaches can be effective. Unlike more general epigenetic marks, they can also be useful for classifying elements based on the particular TFs involved. However, for many biological processes, the TFs involved are not fully known, and so these approaches cannot be applied.

Applications when TFs and TFBSs are known

In early applications, detailed information about TFs involved in muscle determination, such as myogenic factors (MYFs) and Myocyte enhancing factor 2 (MEF2), and their TFBS motifs enabled the prediction of elements that are active in muscle, based on clustering of the TFBS motifs[45,46] (Fig. 1a, Table 1). These and related methods can find up to two-thirds of known muscle enhancers but the validation rate can be low[45]. In *Drosophila melanogaster*, knowledge of the TFs and their cognate TFBS motifs that regulate expression of genes controlling early development enabled several approaches to finding clusters of TFBS motifs relevant to different developmental process[47–51]. All had good sensitivity, in that each found at least one novel enhancer active in transgenic flies, but in most cases the predictions had a low **positive predictive value** (14 to 33%). Modelling matches to TFBS motif matrices as a thermodynamic affinity instead of making binary calls on TFBS motif instances has a substantially higher success rate, probably because many weak matches were able to contribute to the predictions[51]. In this case, only known segmentation genes were investigated; in general, the larger the search domain for predicting CRMs (e.g. whole genome), the lower the positive predictive value.

Applications when TFs and TFBSs are unknown

When relevant TFs and motifs are unknown, motif discovery and CRM discovery can be performed simultaneously. For example, the CisModule software[52] (Table 1) models TFBS motifs and CRMs simultaneously. When applied to the muscle expression dataset described above, this approach recovered some of the known TFBS motifs and showed good specificity in discriminating the true muscle CRMs from random sequences[63]. Training models to discriminate different classes of CRMs (rather than just CRMs from background) can improve the inference of TFBS motifs and CRMs. Smith et al.[53] combine known motifs with motifs discovered to be discriminative between datasets in promoter proximal regions to construct a **logistic regression model** that can significantly predict tissue specific expression in 45 of the 56 human and mouse tissues considered. The ability to discover novel TFBS motifs, especially in the process of CRM identification and classification, will remain important as long as TFBS motifs have not been comprehensively defined.

Future prospects

As collections of TFs and their cognate TFBS motifs are more completely defined, a promising future direction is to build quantitative models that predict expression levels under diverse conditions for both naturally occurring and synthetic CRMs. Impressive success has been achieved for synthetic promoters in yeast using thermodynamic models of

binding affinity of TFs to DNA and to each other[54]. As we strive for an understanding of the regulatory code, experiments such as these will reveal how complete (or lacking) is our knowledge.

Comparative genomics approaches

Assumptions and approaches

Comparative genomic approaches for CRM prediction assume that the DNA sequences involved in gene regulation have remained significantly more similar than non-functional DNA across a wide phylogenetic span, such as multiple species of *Drosophila* or many eutherian mammals. Sequence changes in these regions are thus more likely to show signatures of purifying selection (Fig. 1b). While this assumption holds for most transcription factor coding sequences, it is not uniformly true for CRMs[55,56], as illustrated in **Box 3**. Thus, comparative genomics approaches can be effective only for identifying the subset of CRMs that were under strong purifying selection since the separation of the species under comparison, and they will not reveal lineage-specific, recently evolved CRMs.

Using only signals for evolutionary constraint

Evidence of strong evolutionary constraint in noncoding DNA, without other information such as TFBS motifs, has been used successfully as a *de novo* predictor of CRMs (Table 1). This approach has been applied both at the level of a single TFBS and of an entire CRM.

In alignments of orthologous sequences from a diverse set of mammals, the noncoding regions contain blocks with little or no change among species, surrounded by blocks with sequence differences (Fig. 1b). These conserved blocks are interpreted as functional DNA sequences in which substitutions were rejected during the evolution of the species being compared[57–60]. Noting the similarity between rejection of substitutions in DNA (revealed by the multi-species alignments) and protection of DNA from nucleases by protein binding (biochemical footprinting assays), Tagle et al.[61] called these “phylogenetic footprints” and predicted that they would be reliable indicators of TF binding – even for TFs that have not yet been discovered. This prediction was validated in multiple studies of individual genes and gene families[62–64] (Fig. 1b; Table 1). Subsequently, this approach was part of elegant work to identify regulatory motifs in promoters and 3' untranslated regions of mammalian genomes[65] and entire genomes from multiple *Drosophila* species[66,67]. Because this approach is not dependent on a library of known TFBS motifs, novel motifs are discovered, and these can predict expression patterns of the regulated genes[65] (Table 1).

Evidence of evolutionary constraint over longer segments of noncoding DNA (hundreds of base pairs) can reveal entire CRMs. Early examples are the use of human–mouse alignments to discover enhancers of immunoglobulin[68] and interleukin genes[69]. CRMs predicted by noncoding constraint have been validated as enhancers at a very high rate using reporter gene assays after transfection of cells[70,71] or production of transgenic *Ciona intestinalis*[72], fish (*Fugu rubripes*)[73] or mice[71] (**Box 4**). Hundreds of human noncoding DNA segments showing signatures of extreme evolutionary constraint have been tested for the ability to drive tissue-specific expression in transgenic mouse embryos, and over half were validated[34,74]. In most studies (Table 1), predictions were made in the vicinity of regulated genes[69–72], or a genome was scanned for evidence of extreme evolutionary constraint (e.g. conserved from humans to fish). A much lower validation frequency is observed when these criteria are relaxed[75] (Table 1). Thus many constrained noncoding sequences may not be overtly involved in gene regulation, but constraint combined with other features can be effective for CRM prediction.

From alignments to CRM prediction when TFBS are known

Combining inference of constraint from multispecies alignments with clusters of TFBS motifs can improve CRM prediction. Many known CRMs and *in vivo* bound TFBS motifs were found to be conserved between humans and rodents[15] or among *Drosophila* species[66,67,76], and the specificity of CRM prediction was improved when TFBS motif instances were restricted to those that are conserved in other species[77–79]. Blanchette et al. searched mammalian genomes and alignments for clusters of evolutionarily constrained TFBS motif instances[80]. These predicted regulatory modules (PReMods) encompass a large fraction of known CRMs (Table 1). A subsequent genome-wide mapping of likely enhancers found that over 40% of the DNA segments occupied by the co-activator p300 (which marks many enhancers) overlap with PReMods[81]. Some CRMs are bound by multiple molecules of a TF, each at an individual TFBS, and multiple instances of conserved motifs could represent **homotypic clusters** of TFBSs[82]. When DNA segments with more than one conserved instance of a given motif are tested, they validate at a high rate in transgenic fish and mice (Table 1).

Other efforts focus on specific cell types, e.g. using TFBS motifs for known hematopoietic TFs in addition to multispecies alignments[83]. A limited set of these predictions was tested, and all were validated (Table 1). Recently, Narlikar et al.[84] predicted heart enhancers by applying a model of known and novel TFBS motifs learned from a large set of known heart enhancers to conserved noncoding sequences. This model predicts 42,000 heart enhancers in humans. Of these, 26 were tested in transgenic fish, and an impressive 62% of these were validated.

Although a CRM may be constrained among species, individual TFBS motif instances can tolerate sequence level change[56]. Modelling the evolution of CRMs can capture the signatures of this change without assuming sequence level conservation. The MorphMS model[85] identifies regions in an existing pairwise alignment that fit an evolutionary model derived from a set of existing TFBS motifs, and was found to have the best performance for recovering known *Drosophila* melanogaster CRMs in a comparison of several computational approaches[17]. A promising extension of this approach incorporates gain and loss of binding sites[86] but, due to additional computational complexity, this approach has not yet been employed for genome-wide CRM detection.

From alignments to CRM prediction when TFBS are unknown

Because not all TFBS motifs are known, it is desirable to develop “motif-blind approaches” to prediction that are not limited by current knowledge of TFBS motifs. Approaches that search for patterns in a training set of known CRMs that distinguish them from non-functional DNA have been used for this purpose. One method finds patterns in multi-species alignment columns with significantly more frequent occurrences in training sets of alignments of known CRMs compared to alignments of presumably non-functional DNA[87]. The resulting “regulatory potential” score has been computed across the human and mouse genomes aligned with multiple mammals. Like the approaches based on modelling CRM evolution, this can capture signatures of change rather than just constraint, however using heuristics rather than an explicit evolutionary model. In the vicinity of erythroid-regulated genes, over half of the DNA segments with high regulatory potential that *also* have a preserved match to an erythroid TFBS motif are validated as enhancers in transfected cells[88] (Table 1; an example is *Zfpm1*R13 in **Box 3**). Almost all the PReMods[80] are found in the set of DNA segments with high regulatory potential[89].

A different approach uses multiple methods to search for words (short DNA sequences) that are over-represented in a training set of known CRMs, and then further restricts the word

matches to evolutionarily constrained regions[90]. The predicted CRMs that were tested were all validated both in transgenic *Drosophila* and mice (Table 1).

Advantages, disadvantages, and future prospects

The studies reviewed here illustrate the power of comparative genomics approaches for predicting CRMs, but also highlight substantial differences in validation rate between approaches. The highest validation rates are found when focusing on genes likely to be regulated by a designated set of TFs, e.g. when searching for conserved instances of TFBS motifs for hematopoietic regulators around genes that are expressed in particular blood lineages. Furthermore, studies testing fewer CRMs tend to have higher validation rates. Perhaps it is not surprising that more comprehensive tests that include genes subject to a wider variety of regulatory mechanisms, such as the project examining constrained noncoding sequences on human chromosome 21[75], reveal limited activity of the tested predictions. But the bulk of these studies show partial success of these approaches under favourable circumstances, i.e. involving known TFs and TFBS motifs, and a set of genes responding to a particular stimulus or differentiation pathway.

Some caveats should be kept in mind when evaluating the conservation-based methods. Only a small subset of CRMs is likely to be discovered by extreme evolutionary constraint, e.g. conservation from human to chicken or fish. While this is a strong predictor of developmental enhancers, it does not work equally well in all tissues[74]. Also, perhaps less than 5% of mammalian CRMs show conservation outside eutherian mammals[91].

A major limitation of most comparative approaches is that they are not designed to find CRMs that are active in only one species or that are changing in a lineage-specific manner, such as enhancer GHP88 (**Box 3**). One would expect CRMs that are adaptive for a species to show evidence of rapid evolutionary change, but these will be missed by comparative approaches driven by a search for purifying selection. Indeed, some studies now indicate that most CRMs are species-specific[92].

In future studies, comparative approaches can be developed that cover both closely-related and more divergent species, with the goal of finding lineage-specific and preserved functional sequences, respectively[93]. Additional types of regulatory regions should be tested. A study of silencer and insulator activities for 47 DNA segments from a 1 Mb region containing *CFTR* and flanking genes revealed that signatures of constraint did not improve predictions of these types of regulatory regions[94]. Larger scale studies and the development of models incorporating more types of features could be productive. Also, developing quantitative models that predict expression levels and patterns of target genes, followed by large-scale experimental testing, will be essential to evaluating progress toward more complete understanding of the genomics of gene regulation.

High throughput assays of epigenetic marks

Mapping epigenetic features associated with CRMs

Given the limitations of methods based on sequence motifs and comparative genomics, direct measurement of diagnostic epigenetic features should lead to improved methods for CRM prediction. Epigenetic features are molecules and chemical modifications associated with genomic DNA, including covalent modifications of DNA and histones, RNA transcribed from the DNA, occupancy of DNA by transcription factors, and accessibility of DNA in chromatin to DNases[95]. Particular epigenetic features are highly correlated with CRMs, and progress is being made in finding combinations of these features that may distinguish different types of CRM.

Chromatin immunoprecipitation is a reliable method for purifying DNA in close contact with a particular protein in animal cells, as long as the interactions are relatively stable and a highly specific antibody is available[96,97]. With the introduction of sequence census methods[98], in which the immunoprecipitated DNA is analyzed on massively parallel short-read sequencers, the DNA in close contact with the protein of interest can be determined with remarkable sensitivity and useful resolution (200–300 bp) across an animal genome. This methodology, called ChIP-seq, is being applied in many cell types to find DNA bound by a wide range of transcription factors or associated with chromatin having particular histone modifications (Fig. 1c, **Box 3**). DNase hypersensitivity, a general biochemical feature of CRMs, can also be mapped by sequence census methods called DNase-seq[99,100]. Consortia of multiple laboratories, such as ENCODE[101], modENCODE[102,103], and the NIH Roadmap Epigenomics Mapping Consortium[104] are working in a coordinated manner to expand the coverage of cell types, transcription factors and modifications, and other epigenetic features. This section will summarize advances in using direct epigenetic information to predict two classes of CRM, promoters and enhancers.

Predicting promoters based on TSSs

The TSS is almost invariably located within the promoter (**Box 1**), and promoters can be successfully predicted from the locations of TSSs[105,106]. One study tested 152 predicted promoters by reporter gene assays in a range of mammalian cell types and found that 91% were active in at least one cell type (Table 1, **Box 4**). This remarkably high validation rate, which was confirmed in later studies[106], shows that knowledge of a TSS leads to reliable promoter prediction, with no overt bias for sequence composition or motifs. A different epigenetic feature, the histone modification H3K4me3, is also effective for predicting active promoters[81] (Table 1).

Predicting enhancers based on epigenetic features of CRMs

Enhancers have now been predicted with high accuracy based on several epigenetic features including histone acetylation[107], the histone modification H3K4me1[108], and binding of the co-activator p300 to a DNA segment[109,110] (Table 1). The reporter gene assays were conducted in either transfected cultured cells or in transgenic mouse embryos (**Box 4**). Even when the tests were conducted on large groups of predicted enhancers, tissue-specific expression was driven by 75 to 87% of the DNA fragments, showing that these epigenetic marks are robust, accurate predictors of enhancers. Furthermore, a multivariate **hidden Markov model (HMM)** that combined information on several histone modificationsprovided excellent predictive power for tissue-specific enhancers in human[111] (Table 1).

The success rate of predicting enhancers by occupancy by tissue-specific transcription factors is encouraging but not as successful as using the epigenetic marks just described (Table 1). For instance, of a set of 63 mouse DNA sequences bound *in vivo* by GATA1, half are active as enhancers in transfected cells in culture[112]. As expected from the association of CRMs with evolutionary constraint discussed previously, the set of validated enhancers includes some DNA segments with deep phylogenetic conservation of DNA and conservation of binding between species, but it also includes DNA segments bound in mouse but not human (**Box 3**). A similar fraction of Myoblast determination protein (MYOD)-occupied DNA segments was validated as active enhancers after testing in transfected cells[19], (40%, Table 1, but applying a similar threshold for validation to that used in the GATA1 study shows about half had enhancer activity). Examination of occupancy by multiple transcription factors increases the predictive power of the data. Wilson et al.[113] identified DNA segments jointly occupied by five hematopoietic transcription factors in megakaryocytes. Rather than testing these directly for enhancer

function, they looked for genes previously not known to be important for hematopoiesis that are in the vicinity of the jointly occupied DNA segments. The function of these genes was then tested using a knock-down strategy, and all but one of the knock-downs caused a reduction in blood cell production[113] (Table 1). Thus the jointly occupied DNA segments were excellent predictors of likely enhancers that led immediately to novel insights into hematopoietic regulation.

Advantages, disadvantages, and future prospects

Direct experimental determination of biochemical features associated with promoters and enhancers has many advantages over computational methods. It is grounded in decades of work on biochemical mechanisms of transcriptional regulation, and the proteins and histone modifications being assayed are strongly associated with regulation. The experimental approach is now almost exclusively based on high throughput sequencing and mapping to reference genomes, and while these methods do have some biases, they allow almost complete coverage of animal genomes. These recent advances are exciting, but more research is needed to assess the **sensitivity** and **specificity** of both the comparative and the epigenetic methods. For example, a method based on P300 occupancy predicts hundreds of heart enhancers, whereas almost none of the enhancers predicted by very stringent constraint on noncoding sequences are active in heart[110] (Table 1). However, another approach applying a motif-based classifier to less-stringently constrained sequences predicts 42,000 heart enhancers[84]. Both the latter and the P300-based predictions are validated at impressive rates (Table 1), suggesting that at least some of the current ChIP-seq datasets are missing some CRMs.

One disadvantage of the direct experimental approach is that epigenetic marks must be mapped in tissues and times of development that are informative to the question at hand. Ideally, all transcription factors and all histone modifications would be mapped in all cell types and developmental stages in the species of interest. Achieving this will be difficult for many reasons beyond budgetary ones, such as the limited number of regulatory proteins for which ChIP-quality antibodies are available and the difficulty in obtaining sufficient amounts of many cell types. While the ideal of completeness may never be achieved, a substantial amount of predictive power is likely to be attained as the regulatory landscape is mapped in a large number of cell types. DNase hypersensitive sites are being mapped in a broad range of cell types and tissues[99,100]. These are general marks for sequences potentially involved in regulation; virtually all known CRMs reside in such hypersensitive sites. Some regulatory regions, especially promoters but also some enhancers[81,105,114], are active in multiple cell types. Many others are bound by TFs in specific cell types and can only be identified when assays are done in those cells. The genome-wide maps of epigenetic marks provide a valuable resource for CRM prediction, and one that will increase in value as a broader range of cell types and developmental stages are interrogated.

A caveat in using the genome-wide maps of factor occupancy is that some of these protein–DNA interactions may not be playing an active role in regulation. The very deep coverage achieved in recent ChIP-seq studies reveals significant binding at thousands of sites, but for well-known, lineage-specific transcription factors, the number of bound sites substantially exceeds the number of genes with significant changes in expression in that lineage. For example, over half of all genes are bound by the determination factor MYOD in muscle cells[19]. Integration of information on multiple epigenetic features[111] may allow the TF occupied segments to be partitioned into classes with more specific predicted functions (including no obvious function), thereby giving more accurate predictions.

Future work will also likely interrogate more diverse functions. As noted before, CTCF is almost always found at mammalian insulators with enhancer-blocking activities[115], but it is

currently unclear what fraction of the large numbers of CTCF-occupied segments have this activity. We expect that such surveys will be conducted in the near future.

The bulk of the results summarized in this section were derived from ChIP-seq approaches that yield assignments of TF occupancy at a resolution of 200–300 bp. New technologies should refine that resolution substantially. Already, deep sequencing of DNase-sensitive regions is revealing small segments that correspond to TF binding sites (10–20 bp)[100,116], and a new method employing exonuclease trimming gives very high resolution[117]. Identifying the gene(s) responsive to the TFs at enhancers has been problematic, and many studies used the closest active gene as a proxy for the target gene. However, the high-throughput versions[118–120] of chromosome conformation capture technologies yield three-dimensional interaction maps that are providing exciting new insights into how distal CRMs interact with target promoters.

Recommendations

Comprehensive identification of CRMs is not currently possible from sequence comparisons alone, whether utilized to find clusters of TFBS motifs or to find evidence of strong constraint in DNA. Clusters of TFBSs do not provide sufficient specificity to be used in large-scale investigations, while restricting a search to strong constraint will miss a large number, even a majority, of TF-occupied segments. In contrast, high quality, high throughput biochemical data on epigenetic features will capture a large fraction of CRMs, and of course the fraction captured will increase as the amount of data increases, particularly as more diverse cell types and conditions are assayed. This information is becoming more readily available to individual investigators, either through their own efforts or by using the publicly released data from large consortia. We recommend that this be the starting point in searches for potential regulatory regions, but that both evolutionary information and motif patterns should then be used to bring in insights about potential functions and to organize experimental tests (Box 5).

We expect that future work will show that patterns in the TFBS motifs and their conservation (or lack thereof) can lead to strong and precise functional predictions. At the present time, investigators can use *de novo* motif discovery[121] to predict binding partners of transcription factors and guide further ChIP experiments. It may be productive to partition the TF-occupied segments into motif classes, and assess whether these tend to associate with induction, repression or other activities of likely target genes. Significant associations will probably lead to mechanistic insights.

Perspectives

While most of the evolutionary analysis in the review focuses on purifying selection that has taken place across a wide evolutionary timescale, recent changes in DNA sequences can also affect gene regulation. Some *in vivo* TFBSs are allele-specific[122], e.g. the TF binds to the maternal but not paternal allele in heterozygotes. Genetic variation affecting the affinity of regulatory proteins for CRMs likely explains some of the differences in gene expression between individuals[123,124]. Allele-specific binding by transcription factors or chromatin opening has been found at loci associated with susceptibility to cancer[4,5] and to diabetes[125]. These studies show the impact of recent evolution in CRMs, and point to the medical importance of understanding these recent changes.

Future work should focus on integrating the many types of epigenetic information, building on recent efforts[103,111,126,127]. These could be extended to include multi-species comparisons, not only for the underlying DNA sequences (e.g. to infer evolutionary constraint) but also information about occupancy in additional species. Similarly,

information about *in vitro* binding affinities[128] and motif patterns needs to be brought into the analysis. Such integrations will be challenging, and using them to formulate testable hypotheses will be even more challenging. However, these seem to be reachable goals, and it will be exciting to work toward them. Indeed, the resulting hypotheses will constitute an initial formulation of a possible regulatory code. The hypotheses will need to be tested experimentally, likely starting with conventional gain-of-function reporter gene assays. Larger scale efforts are needed, which require development of higher throughput assays. Also, synthetic biology approaches[54] will provide powerful tests of the hypotheses, and hopefully lead to a better understanding of the regulatory code.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support is from NIH grants R01 DK065806, RC2 HG005573, and U54 HG004695, and funds from Emory University to JT.

Biographies

Ross C. Hardison received his Ph.D. in biochemistry from the University of Iowa and was a postdoctoral fellow at the California Institute of Technology in the laboratory of Dr. Tom Maniatis. He is the T. Ming Chu Professor of Biochemistry and Molecular Biology at the Pennsylvania State University. His current research uses mapping of epigenetic features and comparative genomics to identify *cis*-regulatory modules, their cognate transcription factors, and chromatin states involved in global mechanisms of gene regulation, with a special emphasis on hematopoiesis.

James Taylor received his Ph.D. in Computer Science and Engineering from the Pennsylvania State University and was a postdoctoral fellow at the Courant Institute of Mathematical Sciences at New York University. He is an Assistant Professor of Biology and Mathematics & Computer Science at Emory University in Atlanta, Georgia, USA. His current research in bioinformatics and computational biology focuses on understanding how complex function is encoded in the genome and on making “high-end” computational biology more accessible.

Glossary terms

Purifying selection	The evolutionary process by which deleterious mutations are removed from a population or genome, also referred to as negative selection or evolutionary constraint.
TF binding site (TFBS)	A short segment of DNA that is bound by a particular transcription factor *in vivo*.
TFBS motif	A short string of DNA base pairs (often 6–10bp long) comprising the sequence recognized by the DNA-binding domain of the TF.
TFBS consensus	A string of DNA nucleotides describing the most frequently occurring short sequences in a collection of TFBSs, usually

	including ambiguous positions (e.g. R refers to G or A nucleotides).
Position weight matrix (PWM)	A matrix providing the frequency at which each nucleotide is found at the positions of the TFBS consensus.
Position specific scoring matrix (PSSM)	A matrix providing the log ratio of frequency at which each nucleotide is found at the positions of the TFBS consensus relative to a background model.
Logistic regression	A form of regression used when the output is binary. The predictor is a linear combination of the input variables, transformed with the logistic function to form a probability. For classification, the coefficients are learned to maximize the (log) conditional likelihood of the training data.
Homotypic cluster	A cluster of similar transcription-factor (TF) binding sites, often binding the same TF.
Hidden Markov model	A statistical model in which internal states are not visible but the outputs of these states are, and the outputs can therefore be used to infer the internal states. This model can be used to determine biologically relevant states from ChIP-seq data sets.
False positive	In a prediction experiment, a case where the prediction is positive, but the true class is negative.
True positive	In a prediction experiment, a case where both the prediction and the true class are positive.
Positive predictive value	In a prediction experiment, the proportion of positive predictions that are true positives.
TFBS motif instance	A match to a TFBS consensus or motif matrix within a longer DNA sequence (e.g. a genome or chromosome).
Chromatin immunoprecipitation (ChIP)	A method for purifying the DNA segments in close contact with a TF in living cells. After cross-linking DNA to native proteins in cells and preparing sheared chromatin, antibodies that specifically react with one TF are used to isolate the DNA bound to that TF.
ChIP-seq	A technique for mapping the particular segments of DNA purified by ChIP: it involves massively-parallel short read (second generation) sequencing and then aligning the reads to a reference genome. ChIP-seq is often highly accurate and has very close to whole-genome coverage.
ChIP-exo	An extension of ChIP-seq that includes exonuclease trimming after immunoprecipitation to increase the resolution of the mapped TF bound sites.
Promoter	The DNA sequence that directs RNA polymerase to initiate transcription at the correct place.
Enhancer	A DNA sequence that causes increased expression of its target gene(s).

Insulator	A DNA sequence that controls the ability of an enhancer to regulate a promoter, by an enhancer blocking activity or a domain barrier function, or both.
Position effect	The observation that the level of expression of some genes is affected by their position on chromosomes, with normal level of expression in one location but altered expression when translocated. For example, proximity to centromeres is associated with lowered expression for many genes.
TRANSFAC	A large database of information about transcription factors and their binding sites, including PWMs.
Morpholino oligonucleotide	Synthetic oligonucleotides in which the ribose portion of the nucleotide is replaced a morpholino compound; these are more stable than RNA and can be used to interfere with gene activity in transgenic zebrafish.
Sensitivity	In a prediction experiment, the proportion of the true class that is predicted by the method, i.e. (number of true positives)/(number of true positives + number of false negatives).
Specificity	in a prediction experiment, the proportion of the false class that is not predicted by the method, i.e. (number of true negatives)/(number of true negatives + number of false positives).

References

1. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. Science. 2006; 311:796–800. [PubMed: 16469913]
2. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–116. [PubMed: 1090005]
3. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:9362–9367. [PubMed: 19474294]
4. Tuupanen S, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 2009; 41:885–890. [PubMed: 19561604]
5. Jia L, et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet. 2009; 5:e1000597. [PubMed: 19680443]
6. Harismendy O, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 2011; 470:264–268. [PubMed: 21307941]
7. Farrell JJ, et al. A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. Blood. 2011; 117:4935–4945. [PubMed: 21385855]
8. Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet. 2006; 7:29–59. [PubMed: 16719718]
9. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet. 2008; 9:179–191. [PubMed: 18250624]
10. Rando OJ, Chang HY. Genome-wide views of chromatin structure. Annu Rev Biochem. 2009; 78:245–271. [PubMed: 19317649]
11. Noonan JP, McCallion AS. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet. 2010; 11:1–23. [PubMed: 20438361]

12. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nature Reviews Genetics. 2010; 11:476–486.
13. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012; 13:233–245. [PubMed: 22392219]
14. Frazer KA, Elnitski L, Church D, Dubchak I, Hardison RC. Cross-species sequence comparisons: A review of methods and available resources. Genome Res. 2003; 13:1–12. [PubMed: 12529301]
15. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004; 5:276–287. [PubMed: 15131651]
16. Elnitski L, Jin VX, Farnham PJ, Jones SJ. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res. 2006; 16:1455–1464. [PubMed: 17053094]
17. Su J, Teichmann SA, Down TA. Assessing computational methods of cis-regulatory module prediction. PLoS Comput Biol. 2010; 6:e1001020. [PubMed: 21152003]
18. Zhang Y, et al. Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. Nucleic Acids Res. 2009; 37:7024–7038. [PubMed: 19767611]
19. Cao Y, et al. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell. 2010; 18:662–674. [PubMed: 20412780]
20. Cheng Y, et al. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. Genome Res. 2009; 19:2172–2184. [PubMed: 19887574]
21. Pribnow D. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc. Natl. Acad. Sci., USA. 1975; 72:784–788. [PubMed: 1093168]
22. Carninci P, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006; 38:626–635. [PubMed: 16645617]
23. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981; 27:299–308. [PubMed: 6277502]
24. Fromm M, Berg P. Simian virus 40 early- and late-region promoter functions are enhanced by the 72-base-pair repeat inserted at distant locations and inverted orientations. Mol Cell Biol. 1983; 3:991–999. [PubMed: 6308429]
25. Gillies SD, Morrison SL, Oi VT, Tonegawa S. A tissue-specific transcription enhancer element is locted in the major intron of a rearranged immunoglobulin heavy chain gene. Cell. 1983; 33:717–728. [PubMed: 6409417]
26. Rusche LN, Kirchmaier AL, Rine J. The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae. Annu Rev Biochem. 2003; 72:481–516. [PubMed: 12676793]
27. Martowicz ML, Grass JA, Boyer ME, Guend H, Bresnick EH. Dynamic GATA factor interplay at a multicomponent regulatory region of the GATA-2 locus. J Biol Chem. 2005; 280:1724–1732. [PubMed: 15494394]
28. Jing H, et al. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. Molecular Cell. 2008; 29:232–242. [PubMed: 18243117]
29. Maniatis T, Goodbourn S, Fischer JA. Regulation of inducible and tissue-specific gene expression. Science. 1987; 236:1237–1245. [PubMed: 3296191]
30. Lettice LA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003; 12:1725–1735. [PubMed: 12837695]
31. Schirm S, Jiricny J, Schaffner W. The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. Genes Dev. 1987; 1:65–74. [PubMed: 2828161]
32. Ondek B, Gross L, Herr W. The SV40 enhancer contains two distinct levels of organization. Nature. 1988; 333:40–45. [PubMed: 2834649]
33. Arnosti DN, Barolo S, Levine M, Small S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development. 1996; 122:205–214. [PubMed: 8565831]
34. Pennacchio LA, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006; 444:499–502. [PubMed: 17086198]

35. Landry JR, et al. Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. Blood. 2009; 113:5783–5792. [PubMed: 19171877]
36. Valenzuela L, Kamakaka RT. Chromatin insulators. Annu Rev Genet. 2006; 40:107–138. [PubMed: 16953792]
37. Wallace JA, Felsenfeld G. We gather together: insulators and genome organization. Curr Opin Genet Dev. 2007; 17:400–407. [PubMed: 17913488]
38. Chung JH, Whiteley M, Felsenfeld G. A 5' element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. Cell. 1993; 74:505–514. [PubMed: 8348617]
39. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999; 98:387–396. [PubMed: 10458613]
40. Schoborg TA, Labrador M. The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. J Mol Evol. 2010; 70:74–84. [PubMed: 20024537]
41. Recillas-Targa F, et al. Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. Proc Natl Acad Sci U S A. 2002; 99:6883–6888. [PubMed: 12011446]
42. Huang S, Li X, Yusufzai TM, Qiu Y, Felsenfeld G. USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. Mol Cell Biol. 2007; 27:7991–8002. [PubMed: 17846119]
43. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009; 137:1194–1211. [PubMed: 19563753]
44. Kim TH, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007; 128:1231–1245. [PubMed: 17382889]
45. Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol. 1998; 278:167–181. [PubMed: 9571041]
46. Frith MC, Spouge JL, Hansen U, Weng Z. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. Nucleic Acids Res. 2002; 30:3214–3224. [PubMed: 12136103]
47. Berman BP, 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 U S A. 2002; 99:757–762. [PubMed: 11805330]
48. Markstein M, Markstein P, Markstein V, Levine MS. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci U S A. 2002; 99:763–768. [PubMed: 11752406]
49. Rebeiz M, Reeves NL, Posakony JW. SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. Proc Natl Acad Sci U S A. 2002; 99:9888–9893. [PubMed: 12107285]
50. Halfon MS, Grad Y, Church GM, Michelson AM. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Genome Res. 2002; 12:1019–1028. [PubMed: 12097338]
51. Schroeder MD, et al. Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol. 2004; 2:E271. [PubMed: 15340490]
52. Zhou Q, Wong WH. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A. 2004; 101:12114–12119. [PubMed: 15297614]
53. Smith AD, Sumazin P, Xuan Z, Zhang MQ. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. Proc Natl Acad Sci U S A. 2006; 103:6275–6280. [PubMed: 16606849]
54. Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature. 2009; 457:215–218. [PubMed: 19029883]
55. Chan ET, et al. Conservation of core gene expression in vertebrate tissues. J Biol. 2009; 8:33. [PubMed: 19371447]

56. Ludwig MZ, et al. Functional evolution of a cis-regulatory module. PLoS Biol. 2005; 3:e93. [PubMed: 15757364]
57. Hardison R, Oeltjen J, Miller W. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. Genome Res. 1997; 7:959–966. [PubMed: 9331366]
58. Hardison RC. Conserved noncoding sequences are reliable guides to regulatory elements. Trends in Genetics. 2000; 16:369–372. [PubMed: 10973062]
59. Pennacchio LA, Rubin EM. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet. 2001; 2:100–109. [PubMed: 11253049]
60. Dermitzakis ET, Reymond A, Antonarakis SE. Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet. 2005; 6:151–157. [PubMed: 15716910]
61. Tagle DA, et al. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. 1988; 203:7469–7480.
62. Gumucio DL, et al. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. Mol. Cell. Biol. 1992; 12:4919–4929. [PubMed: 1406669]
63. Hardison R, et al. Comparative analysis of the locus control region of the rabbit beta-like globin gene cluster: HS3 increases transient expression of an embryonic epsilon-globin gene. Nucl. Acids Res. 1993; 21:1265–1272. [PubMed: 8464710]
64. Elnitski L, Miller W, Hardison R. Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the beta-globin locus control region: Role of basic helix-loop-helix proteins. J. Biol. Chem. 1997; 272:369–378. [PubMed: 8995271]
65. Xie X, et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005; 434:338–345. [PubMed: 15735639]
66. Stark A, et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature. 2007; 450:219–232. [PubMed: 17994088]
67. Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. 2007; 17:1919–1931. [PubMed: 17989251]
68. Emorine L, Kuehl M, Weir L, Leder P, Max EE. A conserved sequence in the immunoglobulin Jk-Ck intron: possible enhancer element. Nature. 1983; 304:447–449. [PubMed: 6308460]
69. Loots GG, et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science. 2000; 288:136–140. [PubMed: 10753117]
70. Frazer KA, et al. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. Genome Res. 2004; 14:367–372. [PubMed: 14962988]
71. Grice EA, Rochelle ES, Green ED, Chakravarti A, McCallion AS. Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum Mol Genet. 2005; 14:3837–3845. [PubMed: 16269442]
72. Johnson DS, Davidson B, Brown CD, Smith WC, Sidow A. Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. Genome Res. 2004; 14:2448–2456. [PubMed: 15545496]
73. Woolfe A, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005; 3:e7. [PubMed: 15630479]
74. Visel A, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet. 2008; 40:158–160. [PubMed: 18176564]
75. Attanasio C, et al. Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. Genome Biol. 2008; 9:R168. [PubMed: 19055709]
76. Clark AG, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450:203–218. [PubMed: 17994087]
77. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. 2002; 12:832–839. [PubMed: 11997350]

78. Gibbs RA, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004; 428:493–521. [PubMed: 15057822]

79. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics. 2004; 5:129. [PubMed: 15357878]

80. Blanchette M, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. 2006; 16:656–668. [PubMed: 16606704]

81. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007; 39:311–318. [PubMed: 17277777]

82. Gotea V, et al. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 2010; 20:565–577. [PubMed: 20363979]

83. Donaldson IJ, et al. Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. Hum Mol Genet. 2005; 14:595–601. [PubMed: 15649946]

84. Narlikar L, et al. Genome-wide discovery of human heart enhancers. Genome Res. 2010; 20:381–392. [PubMed: 20075146]

85. Sinha S, He X. MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput Biol. 2007; 3:e216. [PubMed: 17997594]

86. Majoros WH, Ohler U. Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. PLoS Comput Biol. 2010; 6:e1001037. [PubMed: 21187896]

87. Taylor J, et al. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res. 2006; 16:1596–1604. [PubMed: 17053093]

88. Wang H, et al. Experimental validation of predicted mammalian erythroid cis-regulatory modules. Genome Res. 2006; 16:1480–1492. [PubMed: 17038566]

89. Miller W, et al. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 2007; 17:1797–1808. [PubMed: 17984227]

90. Kantorovitz MR, et al. Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. Dev Cell. 2009; 17:568–579. [PubMed: 19853570]

91. King DC, et al. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res. 2007; 17:775–786. [PubMed: 17567996]

92. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010; 328:1036–1040. [PubMed: 20378774]

93. Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. Nat Rev Genet. 2004; 5:456–465. [PubMed: 15153998]

94. Petrykowska H, Vockley C, Elnitski L. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. Genome Res. 2008; 18:1238–1246. [PubMed: 18436892]

95. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell. 2007; 128:635–638. [PubMed: 17320500]

96. Boyd KE, Farnham PJ. Myc versus USF: discrimination at the cad gene is determined by core promoter elements. Mol Cell Biol. 1997; 17:2529–2537. [PubMed: 9111322]

97. Ren B, et al. Genome-wide location and function of DNA binding proteins. Science. 2000; 290:2306–2309. [PubMed: 11125145]

98. Wold B, Myers RM. Sequence census methods for functional genomics. Nat Methods. 2008; 5:19–21. [PubMed: 18165803]

99. Boyle AP, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008; 132:311–322. [PubMed: 18243105]

100. Hesselberth JR, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009; 6:283–289. [PubMed: 19305407]

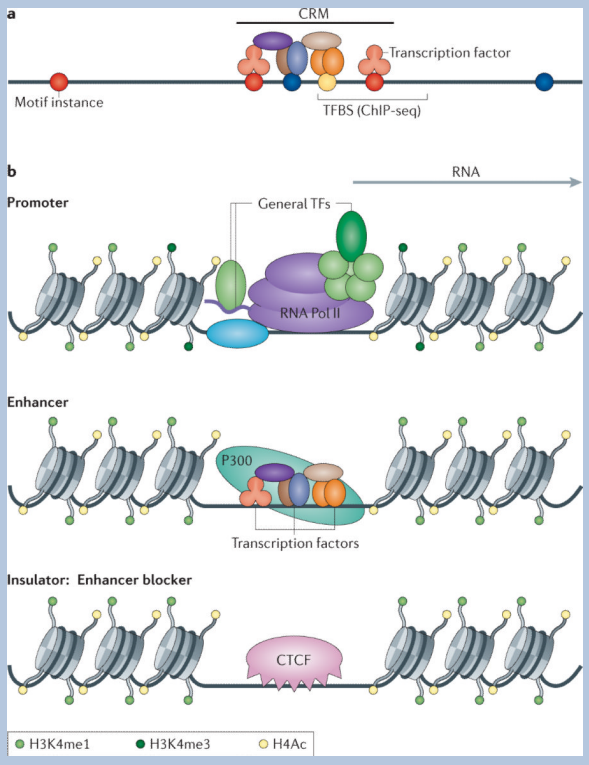
101. ENCODE_Project_Consortium. et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9:e1001046. [PubMed: 21526222]

102. Gerstein MB, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010; 330:1775–1787. [PubMed: 21177976]

103. Roy S, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330:1787–1797. [PubMed: 21177974]
104. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28:1045–1048. [PubMed: 20944595]
105. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM. Identification and functional analysis of human transcriptional promoters. Genome Res. 2003; 13:308–312. [PubMed: 12566409]
106. Landolin JM, et al. Sequence features that drive human promoter function and tissue specificity. Genome Res. 2010; 20:890–898. [PubMed: 20501695]
107. Roh TY, Cuddapah S, Zhao K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev. 2005; 19:542–552. [PubMed: 15706033]
108. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–112. [PubMed: 19295514]
109. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–858. [PubMed: 19212405]
110. Blow MJ, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010; 42:806–810. [PubMed: 20729851]
111. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011:43–49. [PubMed: 21441907]
112. Cheng Y, et al. Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. Genome Res. 2008; 18:1896–1905. [PubMed: 18818370]
113. Wilson NK, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell. 2010; 7:532–544. [PubMed: 20887958]
114. Tuan DY, Solomon WB, London IM, Lee DP. An erythroid-specific, developmental-stage-independent enhancer far upstream of the human “beta-like globin” genes. Proc Natl Acad Sci U S A. 1989; 86:2554–2558. [PubMed: 2704733]
115. West AG, Gaszner M, Felsenfeld G. Insulators: many functions, many mechanisms. Genes Dev. 2002; 16:271–288. [PubMed: 11825869]
116. Boyle AP, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011; 21:456–464. [PubMed: 21106903]
117. Rhee HS, Pugh BF. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell. 2011; 147:1408–1419. [PubMed: 22153082]
118. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2007; 2:988–1002. [PubMed: 17446898]
119. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
120. Li G, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148:84–98. [PubMed: 22265404]
121. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011; 27:1696–1697. [PubMed: 21486936]
122. Kasowski M, et al. Variation in transcription factor binding among humans. Science. 2010; 328:232–235. [PubMed: 20299548]
123. Stranger BE, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–1224. [PubMed: 17873874]
124. Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet. 2009; 10:595–604. [PubMed: 19636342]
125. Gaulton KJ, et al. A map of open chromatin in human pancreatic islets. Nat Genet. 2010; 42:255–259. [PubMed: 20118932]
126. Kharchenko PV, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature. 2011; 471:480–485. [PubMed: 21179089]
127. Hoffman MM, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012

128. Badis G, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324:1720–1723. [PubMed: 19443739]
129. Gilmour DS, Lis JT. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. Proc Natl Acad Sci U S A. 1984; 81:4275–4279. [PubMed: 6379641]
130. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316:1497–1502. [PubMed: 17540862]
131. Robertson G, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007; 4:651–657. [PubMed: 17558387]
132. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
133. He HH, et al. Nucleosome dynamics define transcriptional enhancers. Nat Genet. 2010; 42:343–347. [PubMed: 20208536]
134. Staden R. Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci. 1989; 5:89–96. [PubMed: 2720468]
135. Claverie JM, Audic S. The statistical significance of nucleotide position-weight matrix matches. Comput Appl Biosci. 1996; 12:431–439. [PubMed: 8996792]
136. Schones DE, Smith AD, Zhang MQ. Statistical significance of cis-regulatory modules. BMC Bioinformatics. 2007; 8:19. [PubMed: 17241466]
137. Fujiwara T, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell. 2009; 36:667–681. [PubMed: 19941826]
138. Wu W, et al. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res. 2011; 21:1659–1671. [PubMed: 21795386]
139. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20:110–121. [PubMed: 19858363]
140. Odom DT, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 2007; 39:730–732. [PubMed: 17529977]
141. Cuellar-Partida G, et al. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics. 2012; 28:56–62. [PubMed: 22072382]

Box 1: Features of *cis*-regulatory modules



DNA segments bound by TFs in the nucleus of cells are **TF binding sites (TFBS**, panel a). These are commonly mapped by **chromatin immunoprecipitation** (ChIP)[96,129]. Most *cis*-regulatory modules (CRMs) are comprised of a cluster of TFBSs. A **TFBS motif** is a short sequence (often 6 to 10 bp; colored circle in panel a) found within a TFBS that is required for TF binding, as demonstrated by loss of binding upon mutation of the sequence. The motif can be characterized as a **consensus** or as a position-specific **weight matrix**. Any match to a TFBS motif in a DNA sequence is a **motif instance**.

The size of a TFBS is determined by the resolution of the experimental technique employed. Using chromatin immunoprecipitation followed by high through-put sequencing (ChIP-seq)[130,131], binding of transcription factors *in vivo* can be mapped to a DNA segment about 200–300 bp in length (panel a; note that the TFBS mapped for the bound red motif includes DNA also bound by the orange protein). DNase footprints[100,116] and the recently developed **ChIP-exo**[117] provide higher resolution, approaching that of the bound motif instance.

Different classes of CRM (promoter, enhancer/silencer, insulator) share some chromatin modifications (circles with different shades of green on the blue histone tails extending from nucleosomes, panel b), such as acetylation (Ac) of histones H3 and H4 for all three classes[132] (for simplicity, the Ac is only shown on H4 in the figure) and monomethylation of histone H3K4 (H3K4me1) for both enhancers and insulators (and distal to the TSS around promoters)[108]. Other modifications are distinctive for a CRM class. Active promoters have a nucleosome-depleted region just upstream from the TSS, flanked by nucleosomes with high levels of trimethylation at lysine 4 of histone H3 (H3K4me3)[81,132]. Promoters can also be identified by ChIP-seq for RNA polymerase II. Nucleosomes at enhancers have high levels of H3K4me1[108] and are positioned adjacent to the TFBSs[133]. The co-activator P300 is frequently found at enhancers[108]. Insulators

that work by blocking enhancement require binding by CCCTC-binding protein (CTCF) in mammals[39].

Box 2: Bioinformatic approaches

Bioinformatics approaches for cis-regulatory module identification typically employ supervised machine learning, in which models are built from trusted training data and then used for prediction. Training data are generally derived from experimental data, such as binding footprints or regions identified by ChIP-seq that are enriched for a specific transcription factor, or they are the result of functional assays for enhancer or other CRM functions.

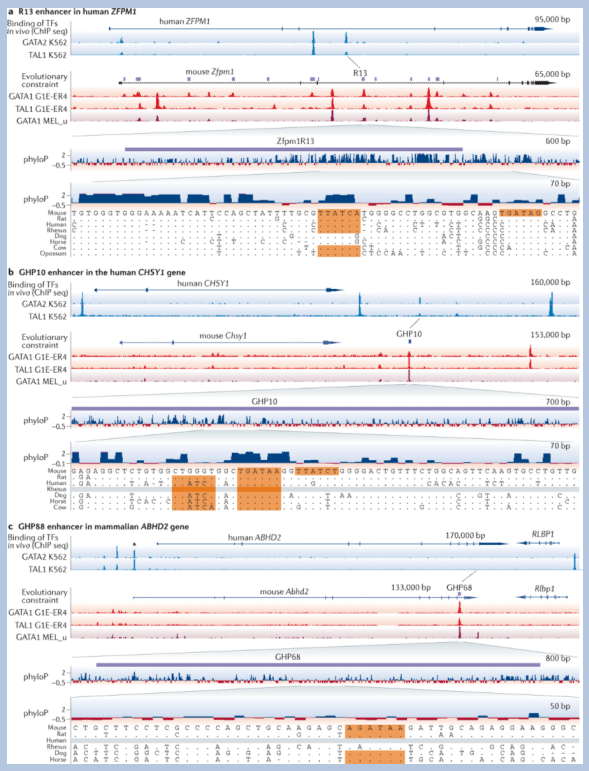
Finding matches to known motifs

TFBS motifs are most often described using a **position weight matrix** (**PWM**), a model for a fixed length sequence that specifies the probability of each nucleotide at each position. Given a background model, a PWM can be converted to a **position specific scoring matrix** (**PSSM**), which can directly compute the log-odds of a given string being generated by the PWM model versus the background model. The log-odds score evaluates a single site, but does not assess the likelihood of finding such a site in a longer sequence. Several approaches can be used to evaluate the statistical significance of these log-odds scores, either through simulating the score distribution[134,135] or from a sequence database[136].

Finding clusters of motif matches

Many approaches have been developed for identifying motif clusters. Simply scanning genomic sequence for windows containing multiple motif matches has been used for predicting CRMs[58], but choosing appropriate significance thresholds can be difficult. One of the first machine learning approaches for CRMs used the positions and scores of strong matches to PWMs as predictors in a logistic regression model. An advantage of such an approach is the ability to capture not just clustering of motifs, but constraints on the organization of motifs in a cluster (such as order). Regardless of the approach used to find clusters, several methods have been developed that use statistical models to assess the significance of motif clusters in a sequence[46], even in the presence of constraints on organization[136].

Box 3: Evolutionary diversity in tissue-specific enhancers

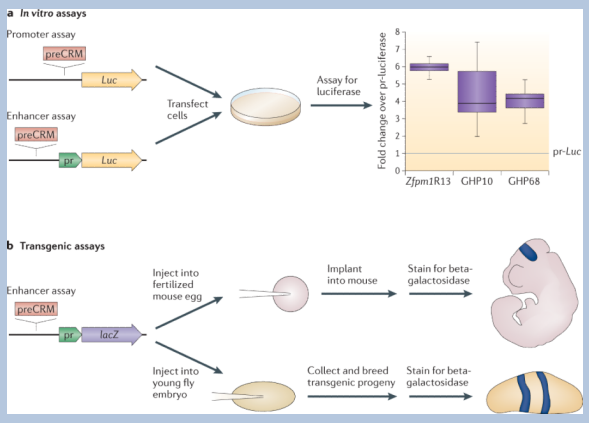


Three examples illustrate that while some enhancers are subject to strong evolutionary constraint over long phylogenetic distances, others show less constraint and still others appear to be lost in a lineage-specific manner. The deeply preserved enhancer R13[88] in the gene *Zfpm1* (panel a) is bound *in vivo* by GATA transcription factors and TAL1 in both human and mouse erythroid cells, as shown by ChIP-seq data[20,137,138], has a strong signature of evolutionary constraint by the phyloP score[139] and contains phylogenetic footprints for a GATA1 binding site motif (boxed in red). The genome browser views are shown at increasing resolution, as appropriate for each feature; known enhancers are shown as blue-filled boxes.

The enhancer GHP10[112] (panel b) is occupied by erythroid TFs in mouse and human, but other predicted CRMs for the *Chsy1* gene differ between mouse and human. GHP10 has a sparser phyloP signal for constraint compared to *Zfpm1R13*, but preservation of some GATA1 binding site motifs.

The enhancer GHP88[112] (panel c) is found in an intron of the mouse *Abdh2* gene, but no GATA1 occupancy is observed at this position in human. In contrast, human-specific binding is seen upstream from *ABHD2* (arrow or asterisk:). While a GATA1 binding site motif is conserved in the rodent, horse and cow homologs of GHP88, no homologous sequence is found in human or rhesus, indicating a primate-specific deletion that leads to a negative signal for phyloP. Gray lines in the alignments indicate that no orthologous sequence is found in the comparison species.

Box 4: Methods for validation of predicted CRMs



The most common methods for demonstrating that a DNA segment can function in the regulation of gene expression are gain-of-function assays after transferring a reporter gene, encoding an readily-assayed enzyme, into cultured cells (transfections, panel a) or whole animals (transgenic assays, panel b). For *promoter* assays (panel a), the predicted CRM is placed in front of a reporter gene (Luciferase, *Luc*) lacking a promoter and transferred into cultured cells. For *enhancer* assays (panel a, lower; panel b), the predicted CRM is added to a reporter gene already driven by a low-activity promoter (pr). The enzyme assays after cell transfection give a quantitative estimate of enhancer activity (panel a, right; box plots show the distribution of enhancement measurements for multiple determinations[88,112]). Information about tissue- and developmental-stage specificity is limited by the cell types investigated by transfection. Staining transgenic mouse or fly embryos carrying the *lacZ* gene encoding beta-galactosidase shows blue staining in the tissues in which an enhancer is active, providing information on tissue specificity. Loss-of-function tests of predicted CRMs (preCRMs), e.g. by targeted deletion, are desirable, but they are more difficult and not used as frequently.

Other methods for investigating predicted CRMs examine the expression patterns of presumptive target genes. The most common assumption is that the gene with a transcription start site closest to the predicted CRM is the likely target. If a likely target gene has an expression pattern expected for the features used to predict CRMs, such as expression in muscle for CRMs predicted by the occurrence of binding site motifs for muscle determination factors, then this supports the validity of the enhancer. Of course, this is not as powerful as a direct experimental demonstration.

A novel approach that uses expression of presumptive target genes is to search for genes not previously known to be required in a tissue of interest. Instead of testing the function of the preCRMs, the effect of specific knock-down of the presumptive target can be monitored, e.g. employing **morpholino oligonucleotides** that interfere with gene function. Defective development or aberrant function of the tissue would serve to validate the activity of the predicted CRMs.

Prediction of binding by a transcription factor to a DNA sequence can be tested by measurement of occupancy *in vivo*, e.g. using chromatin immunoprecipitation. This method is appropriate for determining that a protein is bound to the DNA sequence, but it provides no information about a role in regulation. The older literature contains many studies of binding by purified proteins or proteins in nuclear extracts to specific DNA sequences. Studies with appropriate controls to distinguish specific binding have some

utility, but these results are largely superseded by current ChIP-seq data on *in vivo* occupancy.

Box 5. Steps in prediction and analysis of CRMs

Mapping epigenetic features as the preferred first step

Investigators using genomic data to find transcriptional regulatory regions in animal DNA will find all three approaches to be useful, but each should be employed for a different aspect of their investigations. If data on epigenetic features can be obtained, that should be the starting point for predicting CRMs. High quality datasets on such features provide a relatively unbiased view of the regulatory landscape. We expect that most of the important regulatory regions will be present, assuming the relevant transcription factors are examined in an appropriate cell type for the question of interest. Even if that is not the case, the profile of DNase HSs in a battery of cells across loci of interest could be a good initial guide[101].

Comparative approach can partition candidates

The approaches based on multi-species alignments can then be applied to infer the evolutionary histories of the predicted CRMs and the motifs within them. Indeed, a large number of CRMs may be predicted based on the epigenetic features, and partitioning them based on the extent of phylogenetic conservation can be informative. Conservation can prioritize candidates for functional testing; conservation of TFBS motifs across multiple species of *Drosophila* was found to be strongly associated with regulatory function[66], and GATA1-occupied DNA segments with TFBS motifs that are deeply preserved across mammals were active as enhancers substantially more frequently than those with lineage-specific motifs[112]. However, the hypothesis that evolutionary constraint helps to distinguish TF-occupied segments that are active from bound but passive sites needs much more extensive testing. Most DNA segments bound by a liver transcription factor in one mammal are not bound by that factor at the homologous DNA in a different mammal[92,140], and some lineage-specific occupied segments are active in regulation (Box 3). Thus we recommend using conservation as a means to partition predicted CRMs and to infer their history, but not as a filter to remove them from further consideration.

Partitioning predicted CRMs by depth of conservation may provide insight into the functions of their target genes. An initial exploration of that question found significant enrichments that differed for CRMs conserved to distinct phylogenetic distances[91], and this could be a productive area for more complete investigation. Also, the depth of conservation could reflect variation in the severity of constraint on different aspects of regulatory mechanisms. For example, an interesting hypothesis to test is that CRMs conserved across all vertebrates play a more central mechanistic role in regulation, while lineage-specific CRMs could be modulating that core activity.

Analysis of TFBS motifs for functional prediction

Just as analysis of conservation leads to insights about CRMs predicted by epigenetic features, so will an analysis of TFBS motifs. It is still important to find TFBS motifs for several reasons, including generalizing insights from a well-studied set of CRMs to whole-genome analysis, for making predictions about function, and for understanding the structure of a particular CRM. Epigenetic marks have limited resolution, and motif-based bioinformatics approaches can untangle what is going on inside the modules. Indeed, the conservation analysis just discussed is most informative when applied to the TFBS motifs rather than the entire TF-occupied segment[67,91,112,132]. Furthermore, recent work shows that combining one or more datasets on epigenetic features with TFBS motif models improves the ability to find TF-occupied sites[141].

Online summary

- Predicting *cis*-regulatory modules (CRMs) can both expand our understanding of biology and have applications in medicine and other fields. For example, most of the genetic variants significantly associated with susceptibility to disease are not in protein-coding regions, and we surmise that many affect regulation of gene expression.
- Clusters of TFBSs do not provide sufficient specificity to be used for effective prediction of CRMs in large-scale investigations. However, when applied to identified or likely CRMs, they provide important insights into potentially cooperating transcription factors and help to build a regulatory code.
- Predicting CRMs based on strong constraint in noncoding sequences finds an important subset of the CRMs, that is, those that control developmental regulatory genes. However, they miss a large number, possibly the majority, of TF occupied segments.
- High quality datasets of epigenetic features associated with gene regulation, such as DNase hypersensitive sites, transcription factor occupancy and diagnostic histone modifications, provide a relatively unbiased, sensitive view of the regulatory landscape. The integrative analysis of these features is currently the best approach for predicting CRMs.
- Evolutionary and motif patterns in predicted CRMs should be used to partition the identified regions into categories that could have different functions in regulation.
- Predicting and testing CRMs is essential for deciphering a regulatory code. Synthetic biology approaches, in which DNA sequences inferred for a particular function are synthesized and tested for that function, can assess the accuracy of models for regulatory codes and point to needed improvements.

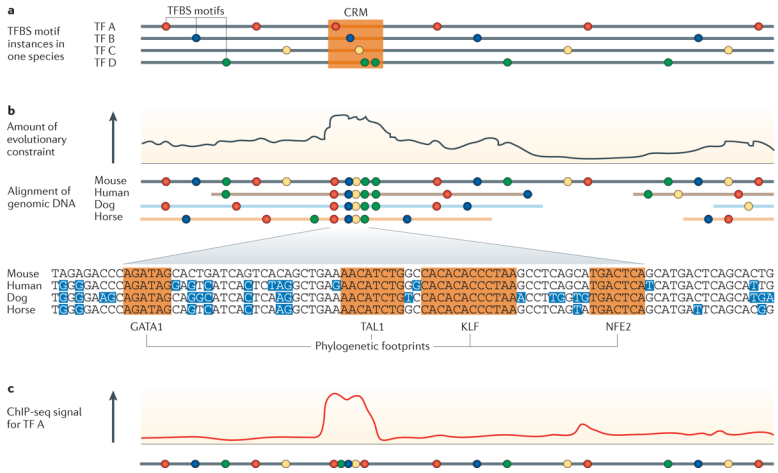


Figure 1. Rationales for three approaches to CRM prediction

(a) Individual instances of TFBS motifs are expected to cluster in CRMs. The instances for each TFBS motif in a sequence (colored circle) are shown on separate gray lines. (b) Evolutionary constraint is expected to preserve clusters of TFBS motif instances in multiple species. All motif instances are shown on a single line for each species. A multispecies sequence alignment shows an example of phylogenetic footprints in a known erythroid enhancer[64]; in the lower panel, nucleotides that differ from the reference sequence (mouse) are colored blue and phylogenetic footprints are indicated by boxes labeled by transcription factors that are known or predicted to bind to them. (c) ChIP-seq assays should show peaks in CRMs. A subset of TF-bound DNA segments will not have a motif instance, which can result from interactions with another bound site.

Nat Rev Genet. Author manuscript; available in PMC 2013 January 10.

Table 1

Reported success rates of different methods for predicting CRMs[a]

Reference	Animal	Biological system	Software or feature	Number of preCRMs	PPV, validation rate	Assay[b]
Clusters of TFBS motifs in a single sequence						
Wasserman and Fickett[45]	Human	muscle	LRA	91	7 of 22 (32%)	Regulated gene proximity
Frith et al.[46]	Human	muscle	COMET	200	4 of 5 (80%)	Regulated gene proximity
Berman et al.[47]	*Drosophila melanogast er]*	Anterior-posterior axis	PATSER, CIS-ANALYST	28	1 of 1 (100%)	Transgenic flies
					10 of 28 (36%)	Regulated gene proximity
Markstein et al.[48]	*Drosophila melanogast er]*	Dorsal-ventral axis	FLY ENHANCER	15	1 of 1 (100%)	Transgenic flies
					5 of 15 (33%)	Regulated gene proximity
Rebeiz et al.[49]	*Drosophila melanogast er]*		SCORE	36	1 of 1 (100%)	Transgenic flies
					7 of 36 (19%)	Regulated gene proximity
Halfon et al.[50]	*Drosophila melanogast er]*	Dorsal mesoderm	ScanACE	647	1 of 7 (14%)	Transgenic flies
Schroeder et al.[51]	*Drosophila melanogast er]*	Segmentation genes	Ahab	52	13 of 16 (81%)	Transgenic flies
Zhou and Wong[52]	Mammal	Muscle	CisModule	29	(54%)	Distinguish positive modules from random
Smith et al.[53]	Human and mouse	Tissue-specific expression	CREAD	1000	45 of 56 (80%)	Tissue specific expression patterns
Phylogenetic footprinting: Strong constraint to identify elements in CRMs or preCRMs						
Gumucio et al.[62]	human	Globin genes	Local alignments	13	12 of 13 (92%)	TF binding in vitro
Hardison et al.[63]	rabbit	Globin gene LCR	yama	3	2 of 2 (100%)	Transfected cells
Elnitski et al.[64]	human	Globin gene LCR	yama	12	1 of 1 (100%)	Transfected cells
Xie et al.[65]	human	All genes	BLASTZ, motif conservatio n score	174	53 of 105 (50%)	Motifs predict expression pattern
Constraint on noncoding sequences						
Loots et al.[69]	human	T-helper cells	PipMaker	90	1 of 1 (100%)	Transgenic mice
Frazer et al.[70]	human	*SIM2*, 21 q	Infer interspecies similarity from hybridization to microarrays	250	10 of 10 (100%)	Transfected cells
Johnson et al.[72]	*Ciona intestinalis*	8 tissue-specific genes	MLAGAN, CHAOS	4	4 of 4 (100%)	Transgenic *Ciona*
Woolfe et al.[73]	*Fugu rubripes*	Regulators of development	megaBLAS T, MLAGAN	1,373	23 of 25 (92%)	Transgenic fish

Reference	Animal	Biological system	Software or feature	Number of preCRMs	PPV, validation rate	Assay[b]
Grice et al.[71]	human	*RET*	AVID, mVISTA	45	15 of 18 (83%)	Transfected cells
Pennacchio et al.[34]	human	developing embryo	BLASTZ	3,100	75 of 167 (45%)	Transgenic mouse embryos
Visel et al.[74]	human	developing embryo	Gumby	2,614	217 of 437 (50%)	Transgenic mouse embryos
Attanasio et al.[75]	human	chr21	PipMaker	2,262	25 of 192 (13%)	Match DNase HSs
					9 of 71 (13%)	Transfected cells
Preservation of clusters of motifs						
Blanchette et al.[80]	human, mouse	Whole genome	PReMod	118,402	236 of 1370 (17%)	TF occupancy in vivo
Donaldson et al.[83]	human	Hematopoiesis	TFBScluster	67	2 of 2 (100%)	Transgenic mouse embryos
Gotea et al.[82]	human	POU3F2	homotypic clusters of TFBS motifs		4 of 8 (50%)	Transgenic fish
		NRF1, E2F4			3 of 3 (100%)	Transgenic mice
Narlikar et al.[84]	human	Heart	Linear regression; enhancer_cl assifier	42,000	16 of 26 (62%)	Transgenic fish
Motif-blind approaches using alignments						
Taylor et al.[87]	human, mouse	Whole genome	ESPERR	282,600		
Wang et al.[88]	mouse	Erythropoiesis	ESPERR+motif match	45,794	24 of 44 (55%)	Transfected cells
Kantorovitz et al.[90]	*Drosophila melanogast er]*	Blastoderm	Enriched words in CRMs	113	5 of 5 (100%)	Transgenic flies
	human	Blood and vasculature		75	2 of 2 (100%)	Transgenic mouse embryos
Biochemical features of promoters						
Trinklein et al.[105]	human	Whole genome	5' end of mRNA	10,276	138 of 152 (91%)	Transfected cells, 8 cell lines
Heintzman et al.[81]	human	HeLa cells	H3K4me1	198	2 of 2 (100%)	Transfected cells
Landolin et al.[106]	Human	Whole genome	5' end of mRNA	37,000	3067 of 4575 (67%)	Transfected cells, 8 cell lines
Biochemical features of enhancers						
Roh et al.[107]	Human	T cells	Histone acetylation; VISTA	46,813	39 of 90 (43%)	Transfected cells
Heintzman et al.[108]	human	HeLa cells	H3K4me17	36,589	7 of 9 (78%)	Transfected cells
Visel et al.[109]	mouse, human	Forebrain, midbrain and limb	p300 occupancy	4,781	75 of 86 (87%)	Transgenic mouse embryos
Blow etal.[110]	mouse, human	heart	p300 occupancy	3,597	97 of 130 (75%)	Transgenic mouse embryos

Reference	Animal	Biological system	Software or feature	Number of preCRMs	PPV, validation rate	Assay[b]
Ernst et al.[111]	Human	9 cell types	Multivariate HMM, integrate histone modifications		8 of 8 (100%)	Transfected cells
Cheng et al.[112]	mouse	G1E-ER4 cells	GATA1 occupancy	63	34 of 61 (52%)	Transfected cells
Cao et al.[19]	mouse	C2C12 muscle cells	MYOD occupancy	25,956	10 of 25 (40%)	Transfected cells
Wilson et al.[113]	mouse	Megakaryopoiesis	joint occupancy	144	8 of 9 (89%)	Engineered knoc-kdowns in fish

A version of this table with additional fields is available as Supplementary Material.

[a]Abbreviations used are *SIM2*, single-minded homolog 2 gene; *RET*, "rearranged during transfection" proto-oncogene; chr21, chromosome 21; POU3F2, POU domain, class 3, transcription factor 2; NRF1, nuclear respiratory factor 1; E2F4, E2 promoter binding factor 4 (E2 is Adenovirus early gene 2); GATA1, GATA-binding factor 1; MYOD, myoblast determination factor; H3K4me1: monomethylation of lysine 4 of histone H3; LRA, logistic regression analysis; COMET, clusters of motifs E-value tool; PATSER and ScanACE, programs that search for matches to TFBS motifs; CIS-ANALYST, FLY ENHANCER, SCORE (Site Clustering Over Random Expectation), programs to find clusters of matches to motif patterns; Ahab, algorithm for predicting CRMs using a thermodynamic model for TF binding; cisModule, an algorithm for inferring CRM locations and TFBS motifs within them; CREAD; a package of programs for motif analysis and CRM prediction and evaluation; yama, PipMaker, MLAGAN, CHAOS, megaBLAST, AVID, VISTA, mVISTA, GUMBY, pairwise and multiple sequence aligners; PreMOD and TFBScluster, pipelines for predicting CRMs based on conservation of clusters of TFBS motifs; enhancer_classifier, program that identifies heart enhancers based on sequence features such as known and putative TF binding specificities; ESPERR, evolutionary and sequence pattern extraction through reduced representations, uses training examples to learn encodings of multispecies alignments into reduced forms tailored to predict CRMs of other functional classes; HS: hypersensitive site.

[b]Most assays monitor the ability of the predicted CRMs (preCRMs) to drive expression of a reporter gene, either in a consistent, tissue-specific manner in the indicated animal ("Transgenic") or at a significantly elevated level in appropriate cell lines ("Transfected cells"). Other assays are the discovery of preCRMs in proximity to genes with predicted regulatory pattern ("Regulated gene proximity") and occupancy *in vivo* by a transcription factor ("TF occupancy").