Genome Biology

REVIEW

Open Access

Decoding enhancer complexity with machine learning and high-throughput discovery

Gabrielle D. Smith[1,2], Wan Hern Ching[1], Paola Cornejo-Páramo[1,2] and Emily S. Wong[1,2*]

*Correspondence: e.wong@victorchang.edu.au

[1] Victor Chang Cardiac Research Institute, 405 Liverpool Street, Darlinghurst, NSW, Australia
[2] School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Kensington, NSW, Australia

Abstract

Enhancers are genomic DNA elements controlling spatiotemporal gene expression. Their flexible organization and functional redundancies make deciphering their sequence-function relationships challenging. This article provides an overview of the current understanding of enhancer organization and evolution, with an emphasis on factors that influence these relationships. Technological advancements, particularly in machine learning and synthetic biology, are discussed in light of how they provide new ways to understand this complexity. Exciting opportunities lie ahead as we continue to unravel the intricacies of enhancer function.

Introduction

Enhancers are a class of genomic *cis*-regulatory elements that play crucial roles in controlling gene expression [1–3]. The term 'enhancer' was first coined in 1981 to describe an element in the simian virus 40 (SV40) genome that enhanced beta-globin gene expression in HeLa cells by 200-fold [4]. We now know that enhancers function in shaping organismal phenotype across all life stages by instructing context-specific transcriptional profiles that vary with cell type, tissue, organ, life stage, and environment [5–7]. Most enhancers reside in non-protein coding regions; however, exonic enhancers have also been shown to drive tissue-specific expression patterns [8].

A significant proportion of mammalian genomes, between 11 and 33%, has been classed as potential enhancers based on genomic associations with markers of enhancer activity across cell and tissue types. This "enhancer real estate" is significantly larger than the 2% of the genome that comprises of protein-coding genes [9–15] (Fig. 1). However, our understanding of the transcription-driving activity of enhancers across cellular contexts remains limited. The vast majority of candidate enhancers have not been validated based on their ability to drive transcription. Large-scale validation approaches have been mainly restricted to in vitro applications and a handful of cell types.

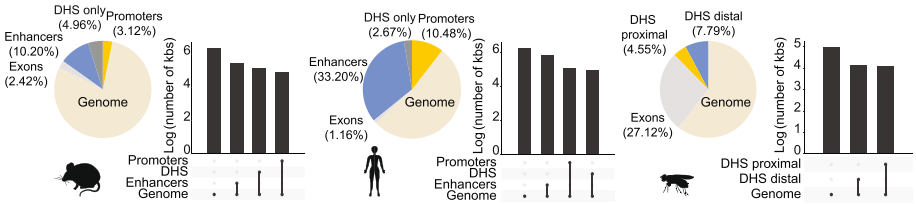

Fig. 1 Proportion of *cis*-regulatory elements in animal genomes. Percentage of the mouse, human, and fruit fly genomes occupied by putative *cis*-regulatory elements based on histone marks, accessible chromatin (from DNase I hypersensitive sites (DHS)), and protein-coding regions (exons; those overlapping predicted regulatory elements are excluded). Upset plots show the $\log_{10}$ number of kbs for each region. Mouse accessible regions were profiled in 55 cell and tissue types and *cis*-regulatory elements were defined by the analysis of H3K4me1, H3K4me3, and H3K27ac histone marks across multiple tissues [15]. Human *cis*-regulatory elements were defined based on 18-state ChromHMM chromatin models across 98 epigenomes [16, 17]; these elements were defined based on their combination of histone modification profiles across the genome. Enhancer and promoter regions were defined as the union of multiple states (EnhWk, EnhA1, EnhG1, EnhBiv, EnhA2 and EnhG2 for enhancers; TssBiv, TssFlnk, TssA, TssFlnkD and TssFlnkU for promoters). Fruit fly accessible regions were profiled in five embryonic stages (S5, S9, S10, S11, and S14) [13], where DHS regions were separated into proximal (±1 kb from TSSs) and distal regions

In recent years, machine learning models trained on epigenomics data has shown remarkable utility in predicting enhancers and their transcription factor (TF) binding sites, including those genetic variants that impact chromatin accessibility [18–24]. For instance, deep learning models have been used to predict the influence of genetic mutations in melanoma by scoring variants that affect chromatin accessibility in melanoma cell states [25, 26]. Computational methods combined with high-throughput synthetic biology have allowed the testing of fully engineered sequences for enhancer activity, broadening our understanding of enhancer evolution and developmental control [27, 28].

Notably, enhancers have been also identified in plants [29] using techniques such as massively parallel reporter assay (MPRA) [30] and chromatin accessibility maps [31, 32]. Plant and animal enhancers appear to have different properties. For example, poised and active animal enhancers are often marked by H3K4me1, but this does not seem the case in plants (reviewed in [33]). This is an important area; however, our focus here will be on metazoan enhancers as they have been most extensively studied.

In this review, we provide an overview of enhancer structure, organization, and mechanisms of action, emphasizing the challenges posed by their rapid evolution and robustness that make classification based on DNA sequences alone difficult [7]. We discuss the use of high throughput, data-rich approaches, particularly in unraveling the "enhancer code" and anticipate that current advancements in molecular biology and computer science will deepen our understanding of sequence-specific enhancer activity, leading to new insights into regulatory mechanisms and evolution. This knowledge will also be essential for incorporating machine learning models in formal disease diagnosis [34, 35].

Mechanisms of action

Enhancers are classically thought to exert their regulatory effects via physical interaction, whereby looping chromatin, supported by structural proteins, brings enhancers, and their associated transcription factors (TFs) into physical proximity with the target promoter [36]. This can bypass linear distances which can span up to a megabase [37–40]. However, the transcription of enhancers, known as eRNAS [41–44], has also

raised questions that the eRNA itself could serve to regulating looping or by forming chromatin domains either locally or in *trans* [45–47].

Enhancer-promoter looping is associated with topologically associating domain-facilitated contact, compatible protein profiles, and distance requirements between the enhancer and target promoter [38, 48]. These features are not universal and other enhancer-promoter communication mechanisms without looping have been described (Fig. 2). Phase separated condensates, a mechanism of biochemical compartmentalization, may enable the action of super enhancers given parallels between the molecular cooperativity in cellular body formation and the assembly of regulatory factors at high density during super enhancer activation [49, 50]. A hypothetical model of communication via diffusion, described as TF activity gradients (TAG), has been proposed to regulate enhancer-promoter communication via short-distance diffusion of acetylated TFs [51]. This model eliminates the need for physical contact between enhancer and promoter and could provide an explanation for observations of proximity, but not contact, between some active enhancers and promoters [52, 53]. Another possible mode of enhancer action could involve the transcription of enhancers into eRNAs, which have been implicated in transcriptional regulation via interaction with NELF, stimulating Pol II pause release and transcriptional elongation eRNAs [54]. Multiple reports of *trans*-acting interactions across homologous chromosomes, a phenomenon known as transvection, has also been characterized in fruit flies [55, 56].

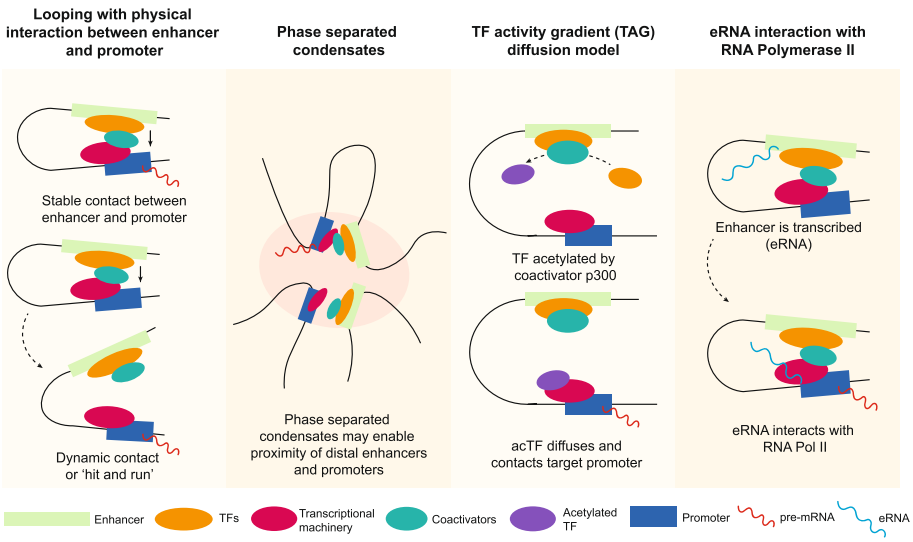

Fig. 2 Mechanisms of enhancer action. Enhancer action on target promoters can occur via looping that enables physical contact in either a stable or dynamic manner in agreement with the evidence of proximity from chromatin conformation experiments and DNA FISH [39]. Alternative mechanism of action include the phase separation of enhancers, promoters, and associated proteins into condensates which can provide the proximity required for enhancer-promoter contact without looping [49, 50]; the diffusion of factors from enhancers to promoters, such as acetylated TFs [51], forming a chemical gradient that enables specific targeting of promoters within close proximity without direct enhancer-promoter contact; and a role for eRNAs by triggering pause release of RNA Pol II at target promoters [54]

Enhancer identification

Enhancers can be identified through various methods, including conservation analysis, genome-wide correlation with chromatin data, measuring eRNAs or transcription of a reporter gene, and using CRISPR-based methods. While sequence conservation can be used to map enhancers from one species to another, many enhancers cannot be identified by sequence conservation alone [5, 7].

The availability of large-scale epigenomics datasets has allowed researchers to analyze genome-wide patterns of regulatory signals to identify enhancers and other genomic elements using correlative approaches. Chromatin enrichment in H3K4me1, H3K27ac, and the chromatin modifier p300 histone acetyltransferase are considered genome-wide markers of regions with enhancer activity [57, 58]. Since enhancers need to accommodate TFs and the associated cofactors necessary for their activation, they are nucleosome deficient. Accessible chromatin away from transcriptional start sites (TSSs), inferred by DNase-seq and ATAC-seq, are also used to detect candidate enhancers [59, 60].

Enhancer RNAs (eRNAs) are bidirectionally transcribed from TSSs within enhancers and tend to overlap known enhancer histone marks. eRNAs can provide higher specificity in enhancer detection compared to histone modifications due to the single base resolution of nascent transcript [61, 62]. eRNAs are identified using assays that enrich for active 5' TSSs, such as CAGE, or nascent transcript assays, such as PRO-seq and GRO-seq, where the expression level of transcripts is considered a functional quantification of enhancer activity [11, 61, 63, 64]. Single-cell transcriptomic profiling can capture the 5′ end of transcripts (CAGE) to identify enhancers at single cell resolution [65]. However, transcription is not exclusive to enhancers but is also a feature of promoters suggesting the regulatory roles of enhancers and promoters are more interchangeable than once thought [62, 66]. Bidirectionally transcribed promoters can act as strong enhancers, while enhancers can also act as weak promoters [63, 67]. These signals provide insights into the role of enhancer as transcriptional hubs and has raised intriguing questions into the biological roles of eRNAs [68]. Beyond the idea that eRNAs are mere passengers of TF activity, some eRNAs have been shown to have specific functions [43], including regulation of spatial organization associated with the production of lncRNA [69], and the formation of transcriptional condensate through m6A methylation of nascent RNAs [70]. Notably, despite significant overlap between the sets of enhancer candidates identified by different approaches, there are incongruencies between the different methods of enhancer annotation [71]. Based on these genome-wide approaches, millions of enhancer candidates have been identified across tissues and cell types in metazoans. However, the validation of these candidates is a significant bottleneck.

In vivo transgenic approaches are used to validate enhancers in a developmental context providing critical spatiotemporal information across the different cell types of a developing animal. These experiments involve the transgenesis of a cassette containing a test sequence with a minimal promoter and a reporter, which may be randomly integrated into the genome or targeted to a safe harbor/neutral landing site using CRISPR/Cas9 [35, 72]. A dual-fluorescence, dual-CRE transgenic cassette can also be used to measure the activities of normal human enhancers and the same enhancer encoding a putative disease variant simultaneously in F1 zebrafish [73]. However, in vivo transgenesis using a reporter gene is low throughput in vertebrates and tend to lack endogenous

context. High throughput validation of enhancer activity, including MPRA and perturbation-based methods will be discussed in a following section.

Enhancer sequence code

Mechanistically, enhancers are considered as clusters of TF binding sites (TFBS) that recruit *trans*-acting factors and target protein-coding gene promoters [74–77]. Parameters including the type, arrangement, and orientation of binding motifs, collectively referred to as enhancer 'grammar' implying a common syntax or logic to the enhancer region, can also play a role in determining enhancer activity (reviewed in [5, 78]). Several models of enhancer organization have been proposed, including the billboard [79], enhanceosome [80], and TF collective models [76]. Each model varies in the mode of DNA-binding protein occupancies and organizational structure (Fig. 3). The enhanceosome model requires the strict arrangement of TF binding sites and direct TF cooperation, while the billboard and TF collective models describe a more flexible arrangement of binding with indirect cooperation—the latter featuring an increased role for protein-protein interactions (reviewed in [2, 5, 78]). Enhancers are thought to fall along a spectrum of these models and the precise mechanisms by which they function can vary depending on the specific enhancer and the cellular context. As such, motif arrangements, mutations or deletions can have varying degrees of impact depending on the

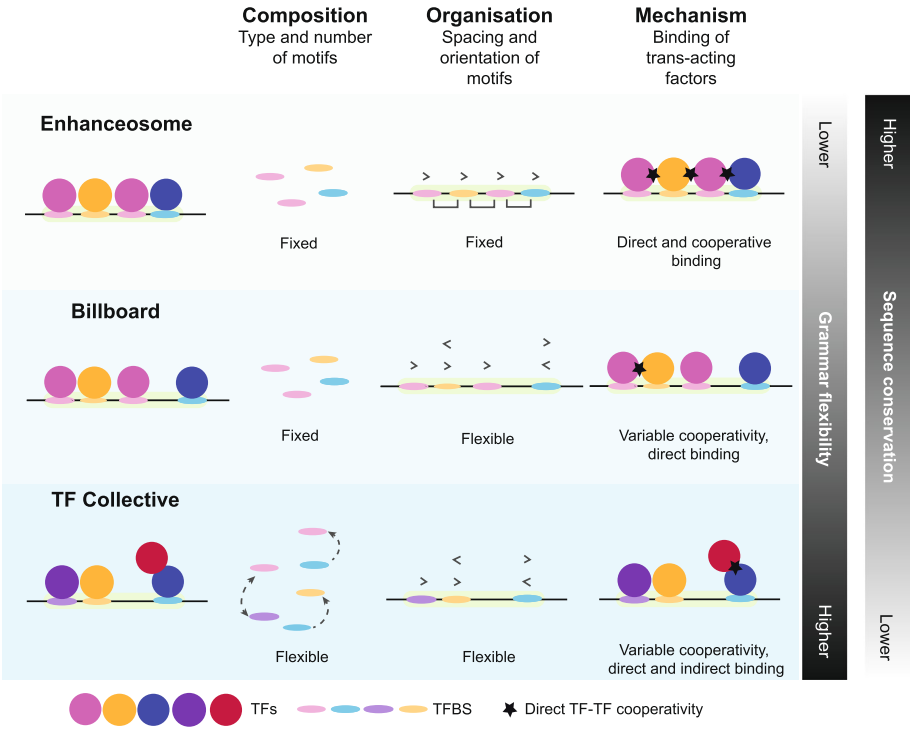

Fig. 3 Current models of enhancer grammar. The flexibility of the type, number, orientation, and spacing of motifs within an enhancer sequence can vary (reviewed in [5, 78]). The enhanceosome model relies on fixed composition, number, and organization of motifs that support direct and cooperative binding of TFs [80], corresponding to enhancers with low sequence flexibility and thus increased sequence conservation. A greater degree of sequence flexibility is found in the billboard [79] and TF collective [76] enhancer models which incorporate variable motif organization and TF cooperativity

enhancer [81, 82]. For example, in the arrangement of ZicL and ETS binding sites that drive gene expression patterns key to *Ciona* notochord development, suboptimal spacing can be tolerated if compensated by stronger TF-DNA binding affinities [83].

An emerging view, supported by studies using transgenic activity assays in vitro and in vivo, suggests that for most enhancers, grammar may be relatively weak and TF occupancy is often sufficient to confer enhancer function [27, 81, 84–87]. It is also important to note that although the major attention is on TF binding motifs, nearby sequences can also impact TF binding by altering DNA shape, chromatin accessibility and allosteric regulation [88–91]. Epigenetic modifications, such as DNA methylation, also interplay with TF binding [92]. While DNA methylation can repress TF binding, some important developmental TFs appear to prefer methylated CpG binding sites [92].

Challenges to understanding enhancer code: evolution

Enhancers, in comparison to promoters, tend to more evolve rapidly and often without strong sequence constraint [7, 93–96]. Positive selection has been observed in a subset of rapidly evolving human enhancers associated with immune function and development [97, 98]. While some enhancers are highly conserved among vertebrates [99–101], around 50% of candidate enhancers detected in 20 placental mammals are lineage-specific and recently evolved [93]. This dynamic reflects a rapid rate of TF binding site turnover and has been linked to transposable elements [102], which make up a significant portion of mammalian genomes [103–105]. Yet, less than half of the lineage-specific enhancers overlap with transposable elements, suggesting that most new enhancers have originated from non-regulatory sequences that already exist in ancestral genomes [93, 103]. In these cases, non-regulatory sequences may have acquired activity through point mutations that create new TFBS [94, 106–110].

Across animal phyla, conserved enhancer sequences are rare. Of five thousand of candidate enhancers from the sea anemone, *Nematostella*, none shared recognizable sequence similarity to *Drosophila* or zebrafish [111]. Only one example of strict sequence conservation extending beyond bilaterians has been reported among animal enhancers [112]. However, around 10% of human-zebrafish syntenic loci, ~ 300, showed conserved TF binding motif arrangements at regulatory regions [113]. Arrangements of TF binding sites have been used to identify many pairs of putative homologous regulatory elements at conserved syntenic loci that otherwise bear little sequence similarity between human and zebrafish genomes [113].

While detecting conserved enhancers across distant metazoan is highly challenging, genome analyses have identified hundreds of examples of microsynteny (pairs of conserved syntenic genes) across metazoans [114, 115]. The long-term linkage of microsyntenic genes across animal evolution is attributed to the presence of a *cis*-regulatory element within a gene regulatory block (GRB) that controls the expression of a developmental gene (the "target" gene) [116, 117]. GRBs with conserved enhancers are found within topologically associating domains that form regulatory, self-interacting chromatin architectural features facilitating long-range enhancer-promoter contacts [118]. In the case of the *Islet-Scaper* microsyntenic region, a sea sponge *Islet* enhancer was able to drive similar GFP expression patterns to those of endogenous zebrafish *Islet* expression, despite the lack of primary sequence similarity [85]. Similarly, teleost enhancers without

detectable evolutionary conservation can direct human gene expression and vice versa [119]. Hence, evolutionary distant animals share similar TFs, TFBSs, and developmental gene regulatory pathways, and enhancer-promoter connections [111, 120–123].

Not all enhancers evolve quickly; some enhancers have stretches of identical sequence that are shared between human, rat, and mouse, and are referred to as "ultraconserved" [99–101]. Ultraconserved elements are often found in large, gene-sparse regions and may represent a subset of a larger group of enhancers that generally have higher levels of sequence conservation and may have substantial differences in their phenotypic contributions [101, 124]. They appear to be characterized by the high occupancy of many TF binding sites [125], which may contribute to their pleiotropy in functional activity between cell types and stages of development, thereby increasing evolutionary sequence constraints [1, 126]. Despite high sequence conservation, mutagenesis at many of these regions does not lead to embryonic lethality, which suggests that these sequences may have negative impacts on fitness at life stages beyond development or are conserved for other unknown reasons [127].

Enhancer conservation also varies between different developmental stages and in different tissue types, although the reasons for this variation are not fully understood [103]. Enhancers defined by ChIP-seq of p300 and open chromatin regions tend to be particularly well conserved at certain critical times during embryogenesis, called the phylotypic stage, when there are similarities in gene expression and body plan within phyla [96, 128]. Cardiac enhancers during mouse embryonic development tend to evolve with less evolutionary sequence constraint compared to forebrain enhancers [95, 96]. Cell-type specific variation may reflect differences in the essential nature of the enhancers or the robustness of the tissues they regulate. Other factors, such as variations in chromatin organization and DNA replication time, may also contribute to the faster evolution of certain enhancers [129].

In summary, our current sequence alignment paradigms appear largely insensitive to *cis*-regulatory conservation. New computational methods based on neural networks is allowing the prediction of tissue-specific enhancers where nucleotide-level conservation is low but the predicted open chromatin in a tissue of interest is conserved [130, 131]. By constraining functional analysis to sequences conserved across great evolutionary distances, we identify only a small proportion of functional information in genomes suggesting new strategies are required.

Challenges to understanding enhancer code: robustness

The resilience of phenotypes to changes in enhancer activity is closely tied to the rapid evolution of enhancer sequences. The effectiveness of natural selection for a phenotype is influenced by its robustness, which refers to the ability of the phenotype to maintain stability in the face of genetic perturbations. Robustness is a general feature of complex systems that are evolvable (reviewed in [132]).

Robust enhancers have a high proportion of genetic sequences that do not impact fitness. These "hidden" variants are expected to evolve neutrally. The robustness of enhancers can be attributed to several characteristics at various organization levels: the structure of individual TF binding motifs, the organization of an individual enhancer, and the arrangement of multiple enhancers within a gene regulatory module (Fig. 4). TF

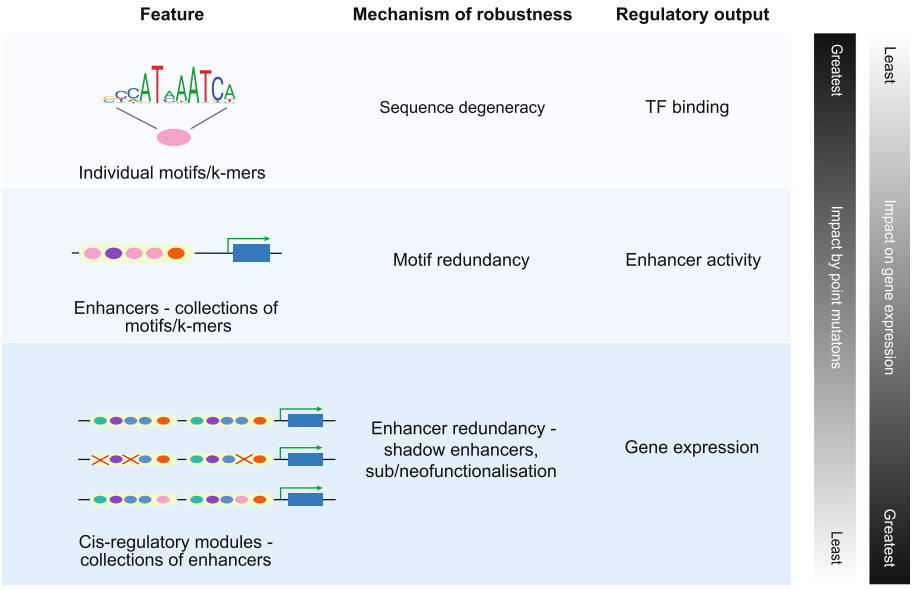

Fig. 4 Organizational structures contribute to robustness and sequence divergence. The short length and sequence degeneracy of TF binding motifs, redundancy and flexibility of motif organization within enhancers, and the structure of the *cis*-regulatory module contribute to the overall robustness of *cis*-regulatory elements

binding motifs contribute to robustness by tolerating base substitutions on a position-specific basis, which is called degeneracy. Another approach to maintain stability in gene activity is by having multiple copies of a motif within enhancers and a flexible motif grammar [133–135]. An example of this robustness is found in *Sepsidae* and *Drosophilidae* flies where the relative position and location of key binding sites that drive the eve stripe 2 enhancers have changed, yet the flies show similar stripe 2 expression patterns [87, 136].

The requirement for low affinity TF binding sites (TFBS) for accurate gene expression patterns during animal development can be also viewed as an emergent property of a robust system [137, 138]. In the *Drosophila* Hedgehog morphogen gradient, low-signaling regions are only active with weak TF affinity [139]. Similarly, the *Ciona* developmental enhancer Otx-a has a "suboptimal" motif sequence and motif arrangement [138]. Low affinity binding contributes robustness because weak binding affinity sites are more likely to randomly occur than strong ones. Most randomly generated TFBS are mutationally distant from the highest affinity sequence [140, 141]. Thus, maintaining a low-affinity binding site is easier than a high-affinity one. Suboptimal binding promotes specificity in gene expression and prevents ectopic expression in non-target tissues, which may have been an emergent trait of a robust system.

There are several factors that contribute to robustness at the level of gene regulatory models. These include through enhancer redundancy [142–144], the need for multiple TFs to bind together [137, 138], and the transmission of genetic signals through different layers of regulatory information [145]. These mechanisms can help maintain the accuracy of gene regulatory circuits despite sequence divergence at *cis*-regulatory elements.

Enhancer redundancy, or the use of multiple redundant enhancers (shadow enhancers) to drive the same gene expression pattern, increases transcriptional robustness

(reviewed in [144]). Shadow enhancers regulate the expression of the same gene, compensating for environmental or genetic alterations to normal developmental programming [143, 146–150]. Many shadow enhancers are partially functionally redundant, with enough overlapping spatial activity maintaining robust developmental gene expression and buffering the impact of genetic variations [150]. Genes with greater regulatory complexity, including more shadow enhancers, results in more robust in gene expression by comparing *cis*- and *trans*-acting genetic variation in *Drosophila* F1 lines [145].

The binding of multiple TFs functions similar to logic gates, masking the impact of mutations and increasing the accuracy of transcriptional control [137, 138, 151–153]. Propagation of genetic signals through multiple regulatory layers helps to maintain the fidelity of gene expression patterns [145, 154, 155]. Thresholds on transcriptional activation or repression can buffer signal variation.

The interplay between evolvability and robustness is a recurring theme in the study of animal regulatory networks. Robustness can promote diversity, leading to the increased evolvability of phenotypes. The short length of TF binding sites allows new TF binding sites to emerge quickly during evolution [156], enabling even random sequences to acquire *cis*-regulatory activities [28, 108]. For example, it takes 0.5–10 million years to evolve the complexity required for a *cis*-regulatory element involved in anterior-posterior axis specification in *Drosophila* blastoderm, starting from a random genome background [107]. In a study using mutational libraries in *Drosophila* embryos, Galupa et al. showed that while existing developmental enhancers are constrained in cell-type specific function, de novo elements harboring TF motifs can drive developmental gene expression across different cell types [28]. Increased levels of sequence variation at developmental enhancers may have propelled speciation and morphological diversity [97]. An experimental evolution study in *E. coli* show that new mutations can become quickly fixed in the population, even in the absence of selection [157].

The concepts of neofunctionalization and subfunctionalization, proposed by Ohno [158] to explain the fate of duplicated genes and the emergence of new functions, also apply to the evolution of duplicated enhancers. Redundancy of function in shadow enhancers can contribute to new gene regulatory networks [142]. The pace of enhancer turnover and larger number of enhancers suggest that these processes occur more frequently in enhancers than in genes.

The mode of TF binding affinity inheritance can also enhance regulatory evolvability. Unlike gene expression, which is often inherited in a dominant or recessive manner, TF binding occupancy at *cis*-regulatory elements typically follows a co-dominant inheritance pattern. This may allow genetic variants that contribute to regulatory differences to be easily selected for, promoting adaptability in gene regulatory networks [6, 145, 159].

Investigating enhancer activity by high throughput experimentation

Experimental validation of enhancers is necessary to confirm enhancer activity and understand the relationship between enhancer sequence and function. This poses a significant challenge due to the context-specific nature of enhancers and the sheer number of enhancer candidates. Validation of enhancer activity can be performed using

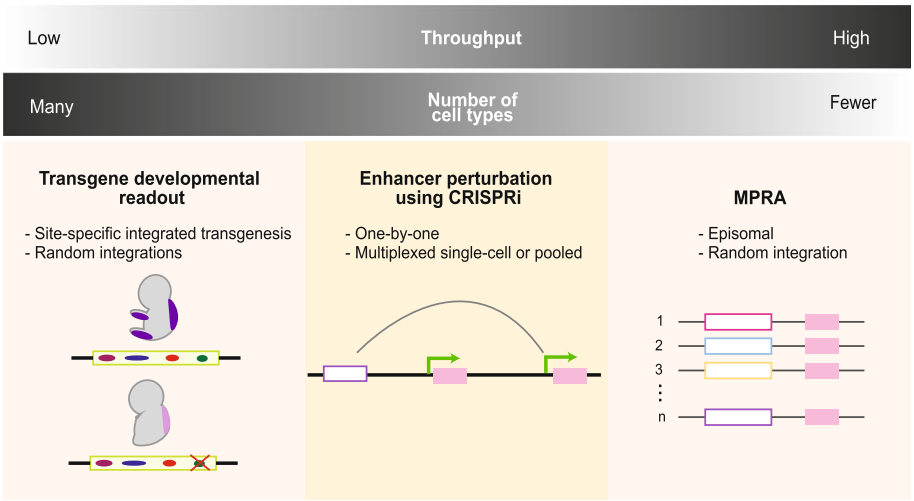

Fig. 5 Experimental methods for testing enhancer activity. Methods that are used to assess enhancer activity involve a trade-off between the number of sequences that can be tested and the number of cells assessed at one time. Developmental transgenic approaches can reveal enhancer activity across many cells at the same time on a per sequence basis. On the other hand, massively parallel reporter assays (MPRA) are able to assay thousands of sequences by random integration or in an episomal manner. Perturbation experiments using CRISPR interference (CRISPRi) can reveal transcriptional targets and can be combined with single-cell readouts to increase throughput

transgenic animal models or in high throughput using massively parallel reporter assays (MPRA) and CRISPR-based perturbations (reviewed in [160, 161]) (Fig. 5).

MPRAs employ a library of reporters and high-throughput sequencing to examine potential enhancers [27, 81, 86, 162, 163]. MPRAs can simultaneously assess thousands of potential enhancer sequences, using gene expression as an indicator of enhancer activity [135, 164]. In this approach, a library containing thousands of plasmids, each carrying an enhancer sequence adjacent to a minimal promoter, is introduced into cells or animal models. MPRA libraries may be randomly integrated into the genome allowing the study of chromatin location-specific effects, or remain separate from the genome (episomal), reflecting the overall regulatory capacity in the tested cell type [165–168].

The design of MPRAs, including factors such as oligo length, the relative positioning of the candidate sequence to the promoter, and the choice between integrated or episomal assays, can influence reporter activity. A comparison of nine major strategies by Klein et al. showed that while most MPRA designs correlate well, the location of the enhancer candidate on the plasmid has a more significant impact than the differences between episomal versus integrated assays [169]. Additionally, while sequence orientation generally does not matter, sequence length, which influences the number and type of binding sites present, can strongly influence activity outcomes.

MPRAs have enabled researchers to validate the activity of endogenous *cis*-regulatory elements [135, 170] while facilitating investigations on the impact of human genetic variations [171, 172]. Studies have varied the positioning, orientation, and diversity of TFBS for key pluripotency factors in stem cells revealing that motif grammar is often flexible but mutations within TFBS can disrupt binding and affect activity [135, 164, 172–174].

Limitations to MPRAs include a lack of endogenous chromatin context, and a loss of relevant epigenetic modifications. As MPRAs are typically used in homogenous cell populations, this restricts their application in rare cell types or cells that are challenging to maintain in culture. However, recent advancements have enabled MPRAs to be combined with single cell RNA-seq sequencing, allowing researchers to study enhancers during cell differentiation and paving the way for the evaluation of enhancers in their native cellular contexts [175, 176].

CRISPR-based genetic perturbation screening addresses the limitations of MPRAs by studying enhancers in their natural cellular context. This technique can be applied on a large scale and at single-cell resolution, enabling the investigation of multiple loci by introducing various perturbations to many cells. Activation or repression of regulatory elements can be examined using CRISPR interference (CRISPRi) or CRISPR activation (CRISPRa) or by direct editing of the regulatory sequence. Using pooled guide RNAs and high-content readouts, these methods allow for the determination of direct and indirect relationships between enhancers and genes at multiple *cis*-regulatory elements [177–186]. Although these techniques are usually performed on cells in vitro, there are in vivo applications using adeno-associated viral (AAV) in animal studies [187].

Using machine learning to dissect *cis*-regulatory elements

Machine learning is transforming our understanding of cis regulatory sequences and their role in gene regulation [188] (Table 1). By using large datasets of multi-omics information, or data from MPRA experiments, deep learning algorithms can identify complex patterns and relationships within the data that may be difficult to otherwise discern. The flexibility of these algorithms has seen them applied to a range of challenging problems. For example, to differentiate between all mapped human *cis*-regulatory elements [18], identify cell-type specific accessible chromatin [20], predict TF binding sites and enhancers across species [189–191], prioritize the impact of regulatory mutations [192], dissect enhancer and promoter grammar [27, 193], and to predict gene expression [163, 194, 195].

A general usage example is as follows: a machine learning algorithm is trained on a pre-defined set of features, such as publicly available datasets of functionally validated enhancer sequences, histone markers, and open chromatin, by associating the input data with labels. The algorithm is then able to determine the underlying patterns that contribute to the labeled class. This process, called training, involves minimizing a loss function (e.g., classification error) at each iteration of the algorithm. The training set typically consists of a fraction of the total available dataset, while the test set is a held-out subset used for model evaluation and is not used in training. A diverse training dataset can improve prediction accuracy and reduce bias in the model [197]. Models can also be trained on data from specific biological contexts and then used for inference in different contexts [189–191, 199, 209]. For example, a model trained to distinguish enhancers in one species can be used to infer enhancers in another [189–191]. Training is typically the most time-consuming and memory-intensive part of machine learning and often requires specialized hardware such as GPUs (graphics processing units).

As input, many studies have leveraged large-scale epigenomics datasets from global consortium initiatives, such as the human and mouse ENCODE and NIH Roadmap Epigenomics Consortium projects, which comprise multiple omics readouts across a wide range of cell lines and primary tissues. The Cistrome Data Browser is a useful resource that compiles all publicly available human and mouse ChIP-seq and DNase-seq datasets [210]. The candidate enhancers from primary tissue data have typically not been experimentally tested for enhancer activity. However, some studies have used experimentally validated enhancers, such as the enhancer VISTA database [211], to train machine learning models to identify tissue-specific enhancer syntax [203]. Sequence models trained with activity data from MPRA experiments can be used to identify the sequence basis for regulatory activity [27, 163, 193].

A multitude of machine learning algorithms have now been developed for regulatory element prediction, with neural network frameworks becoming increasingly popular (Table 2). Non-neural network algorithms comprise of a range of machine learning methods, including support vector machines (SVMs), and tree-based approaches such as random forests (RFs) and gradient boosting machines (GBMs) (Table 1).

Enhancers can be represented as position weighted matrices (PWMs) derived from validated TF binding sites, as k-mers, or using one-hot encoding (reviewed in [7]). One-hot encoding is a method that converts each nucleotide to a numeric variable and commonly used in neural network models. PWMs are easily interpretable but are limited to the motifs of selected proteins. K-mers and gapped k-mers are more flexible representations because they capture all combinations of short sequence patterns, allowing for the de novo discovery of motifs. The gapped k-mer support vector machine (gkmSVM) approach has consistently outperformed its predecessor, kmer-SVM, and has been widely used to analyze enhancer sequences [23, 190, 196]. The most predictive k-mers from these models often match known experimentally confirmed TF binding motifs [21]. The impact of regulatory variants can be assessed by calculating the differences in gkmSVM scores, termed deltaSVM [24]. While gkmSVM is effective and easily interpretable, it may not be able to recognize long-range patterns between motifs due to cooperative or additive TF binding.

Over the past decade, convolutional neural network (CNN) has emerged as a powerful neural network architecture. The complex interconnected multi-layered neuron structure in neural networks allows the algorithm to discern patterns and features that may not be otherwise recognizable [219]. To increase the capacity for such complex pattern recognition, there can be many layers of neurons in these networks, leading to the term "deep learning." Convolution refers to the use of a filter window of a certain length to smooth out noise while retaining important features.

CNNs can be used alone and as part of hybrid frameworks [219]. Early applications of CNNs to genomic data include CSI-ANN [204], DeepBind [200], DeepSEA [192], and Basset [20] (Table 1). These were trained to predict TF motifs, prioritize functional variants at regulatory regions, and classify features such as chromatin accessibility from the sequence. These methods laid the foundations for other high-performing methods designed for regulatory elements, such as DanQ [201], DeepEnhancer [199], DeepMEL [191], and DeepSTARR [193] (Table 1).

Table 1 Machine learning models used in the prediction of *cis*-regulatory elements

Method	Core algorithm/architecture	Goal	Trained model	Reference
Gkm-SVM	Support Vector Machine	To find distinguishing features within regulatory elements	**Class 1**: CTCF ChIP-seq signal enriched regions in GM12878 cell line **Class 2**: Random sequences (matching length, GC and repeat fraction)	[196]
EnhancerFinder	(Multiple Kernel Learning) Support Vector Machine	Enhancer prediction (developmental enhancers)	**Class 1**: Enhancers from VISTA Enhancer Browser **Class 2**: Random regions from genomic background	[197]
RFECS	Random Forest	Enhancer prediction	**Class 1**: p300-binding sites (H1 and IMR90 datasets from NIH Roadmap Epigenome Project) **Class 2**: TSS overlapping DNase-I, and random regions distal to known TSS and p300 sites (H1 and IMR90 datasets from NIH Roadmap Epigenome Project)	[198]
DeepEnhancer	Convolutional Neural Network	Enhancer prediction	**Class 1**: Enhancers from FANTOM5 **Class 2**: Sequences from human reference genome	[199]
DeepSEA	Convolutional Neural Network	To prioritize functional variants at regulatory regions	**Multi-label:** Open chromatin, TF binding and histone mark profiles from ENCODE and Roadmap Epigenomics datasets across multiple human cell types	[192]
DeepBind	Convolutional Neural Network	TF binding prediction	**Class 1**: Protein binding microarrays, ENCODE ChIP-seq peaks, HT-SELEX **Class 2**: Shuffled class 1 sequences (maintaining dinucleotide composition)	[200]
Basset	Convolutional Neural Network	To find distinguishing features within regulatory elements	**Multi-label:** Chromatin accessibility in 164 cell types (ENCODE and Roadmap Epigenomics Consortium)	[20]
DeepSTARR	Convolutional Neural Network	To find distinguishing features within regulatory elements	**Class 1**: Enhancers with developmental activities **Class 2**: Enhancers with housekeeping activities	[193]
BiRen	Convolutional Neural Network + (Gated Recurrent Unit) Bidirectional Recurrent Neural Network	Enhancer prediction	**Class 1**: Human and mouse enhancers from VISTA Enhancer Browser *with* reproducible expression patterns **Class 2**: Human and mouse enhancers from VISTA Enhancer Browser *without* reproducible expression patterns	[22]
DeepMEL	Convolutional Neural Network + (Long-Short Term Memory) Bidirectional Recurrent Neural Network	To find distinguishing features within regulatory elements	**Multi-label:** Melanoma human open chromatin regulatory regions	[26]

Table 1 (continued)

Method	Core algorithm/architecture	Goal	Trained model	Reference
DanQ	Convolutional Neural Network + (Long-Short Term Memory) Bidirectional Recurrent Neural Network	To find distinguishing features within regulatory elements; To prioritize functional variants at regulatory regions	**Multi-label:** 919 ChIP-seq and DNase-seq peaks from ENCODE and Roadmap	[201]
AgentBind	Convolutional Neural Network	Predicting TF binding sites	**Class 1**: ENCODE TF binding ChIP-seq data from multiple cell types **Class 2**: Genome-wide excluding Class 1 regions matched for GC content	[202]
ResNets	(Residual Network) Convolutional Neural Network	To find distinguishing features within regulatory elements	**Multi-label:** Enhancer sequences with distinct regulatory architectures (homotypic clusters, heterotypic clusters, enhanceosomes)	[203]
CSI-ANN	Time-Delay Neural Network	Enhancer prediction	**Class 1**: HeLa cell ENCODE data, Human $CD4^{+}$T cell data **Class 2**: Random genomic loci	[204]
EnhancerDBN	Restricted Boltzmann Machine + Deep Belief Network	Enhancer prediction	**Class 1**: Human "positive" enhancers (VISTA Enhancer Browser), DNA methylation, histone marks, GC content **Class 2**: Genomic background matched for length and chromosome distribution to **Class 1**	[205]
BPNet	(Residual Network) Convolutional Neural Network	To predict TF binding profiles at single base-resolution	**Multi-label:** ChIP profiles for TFs	[206]
DNABERT	Bidirectional Encoder Representations from Transformers	To find distinguishing features within regulatory elements	**Multi-label:** k-mers	[207]
Sei	Convolutional Neural Network; linear and non-linear layers with residual connections	Classifies based on > 21,000 types of human chromatin profiles	**Multi-label:** > 21,000 types of publicly available human chromatin profiles (TF binding, histone marks and DNA accessibility) across > 1,300 human cell lines and tissues	[18]
Enformer	Convolutional Neural Network + Transformer	To predict gene expression and chromatin state profile across multiple cell types in human and mouse genomes	**Multi-label:** 5,313 human and 1,643 mouse gene expression and chromatin states at 128 bp resolution from 200 kb of input sequence	[194]
ChromBPnet	Convolutional Neural Network	To predict chromatin accessible profiles at single base-resolution across the genome after removing biases from enzymes used in DNase-seq and ATAC-seq assays	**Multi-label:** SnATAC-seq of human developing cortex	[208]

Table 2 Common architectures for *cis*-regulatory classification

Machine learning algorithm	Mechanism	Advantages	Interpretation
Support vector machine	Finds a maximal margin hyperplane that best divides data into the required classes	Relatively memory efficient and best suited for high numbers of input dimensions (e.g., k-mers)	With respect to GkmSVN: - Calculation of importance scores at nucleotide resolution using Shapely values, GkmExplain [212] - Introduction of variants in the input sequence and estimation of their impact on the SVM score, deltaSVM [24]
Random forest	Predictions are made from the aggregated result from a set of decision trees, trained in parallel, where each node represents a particular feature	Features are used as explicit classifiers, providing a easy way to interpret the model	- Estimation of feature importance scores, such as Gini score, permutation score, and Shapley values, is a standard practice for dissecting tree ensembles [191, 213] - Partial dependence plots are useful to interpret a random forest; they show the relationship between a given feature and the response variable while other predictor features remain constant [214]
Gradient boosting machine	Uses a series of random forests, and allows for the systematic decrease of a loss function with forests improving on one after another	Yields the benefits of random forests but with added robustness due to having continually improving forests	Similar to random forest
Convolutional neural network (CNN)	Filters of varying sizes slide across the sequence/input unit, capturing patterns and integrating information using cross-correlation to produce a feature map of the sequence	Can learn complex patterns while reducing dimensionality compared to non-convolutional neural networks	Reviewed here [215] - Search for subsequences that activate a convolutional filter and construct PWMs - Attention weights for visualizing feature importance - Propagation of perturbed data through model to observe effects on predictions. This can be done by forward propagation (in silico mutagenesis (ISM)) or backward propagation (e.g., GradCAM, DeepLIFT [216]) - Aggregation of attribution maps to identify globally important sequence motifs (e.g., TFMoDisco [217]) - Initializing filters to known TF motifs (e.g., DanQ [201])
Bidirectional recurrent neural network (RNN) Related: Time-delay neural network	Hidden states in layers preserve information from previous layers, forming a context that contributes to deciding the next action	Captures interdependencies between hidden states	Similar to CNN
Bidirectional Encoder Representations from Transformers (BERT)	Uses an attention-based model used in natural language processing (NLP) tasks	Use self-attention to understand interaction between important regions	- Shapley values can be computed to dissect BERT models [218] - DNABERT-viz was developed to visualize importance scores at nucleotide resolution leveraging self-attention values [207]

Model performance, measured by the area under the curve comparing false positive versus true positive rates (ROC-AUC), exceeded 80% in many chromatin feature classification tasks, such as distinguishing between cell types. However, this metric may convey an overly optimistic impression of these models' performance in cell-type classification tasks due to significant class imbalance.

Natural language processing (NLP) models, such as GTP, have achieved impressive capabilities in different tasks and could surpass CNN-only models in detecting distant semantic dependencies within genetic sequences. Large language models may excel at discerning complex dependencies between sequence elements [220] (Table 2). For example, BERT (Bidirectional Encoder Representations from Transformers) [220] has achieved state-of-the-art performance in NLP tasks and holds promise for improving our understanding of the genome. DNABERT, a BERT model pretrained on the human genome using k-mers as inputs, has developed a general-purpose understanding of the genomic semantics and has been applied to classify promoters and identify TFBS [207].

Another transformer-based model, Enformer, is a large deep learning algorithm trained on ~ 7000 human and mouse datasets, that has shown high performance in predicting cell-type accessible chromatin and gene expression across human and mouse genomes [194]. Karollus et al. showed that Enformer has learnt the causal principles of key TFBS at promoters in K562 cells but that it does not sufficiently account for distal enhancer activity [221]. This is likely due to class imbalance as the number of enhancers driving a target gene's expression decreases with distance away from the gene's TSS. These findings underscore the importance of conducting further research to determine whether deep learning sequence-based models employ correlative or causal sequence principles in their predictions.

While deep learning algorithms can make accurate predictions, they can also be difficult to interpret and are often referred to as "black box" algorithms due to their lack of transparency [222]. The interpretation of AI models is an area of ongoing development in genomics research (reviewed in [215]). The architecture of a neural network can influence its interpretability, with designs that tend to learn either distributed (partial) or localist (whole) representations of sequence motifs with the latter providing a greater level of insight into network decisions [223]. Several methods have been developed to assign importance scores to individual nucleotides to interpret deep learning models. These include DeepLIFT [224], which uses a difference-from-reference method, and DeepExplainer, which uses Shapley Values [216]. Shapley Value is a concept from game theory that considers the contribution of each feature not just based on its input order but also in all other possible orders, to provide a fair assessment of each feature's importance. Another method, called TF-MoDISco [217], is specifically designed for motif interpretation and discovery and is able to process sequence importance scores using information from all the neurons of a neural network. This method can also be used with feature attribution importance scores from gapped k-mer support vector machines (GkmExplain) [212]. Clustering algorithms are used in the interpretation of machine learning frameworks to identify important motifs which are then compared to PWMs. The clustering of motifs is key to the interpretation of trained models [191, 217].

Model interpretation can be facilitated by using simple network architectures. For instance, ExplaiNN [225] uses a large series of simple neural networks each of which learns different TF binding profiles making it efficient to train and allowing for global interpretability while sacrificing the ability to capture interactions between different motifs.

Other approaches to interpreting decision-making processes in neural networks include modifying the input data to test the importance of specific nucleotides and analyzing the network structure (for a detailed review, see [180]). Studies have shown that up to 50% of motifs learnt using different machine learning methods do not match any known canonical TFBS. This may be due to algorithmic limitations or that these motifs may have biological roles other than protein recognition.

Challenges and opportunities

The field of machine learning is rapidly evolving, with models demonstrating great potential in their ability to identify enhancer sequences. Sequence models have the potential to play important roles in prioritizing disease-causing variants and in defining cell-type resolved *cis*-regulatory elements when combined with MPRA and single cell genomics (e.g., congenital heart disease [226]). Despite these exciting developments, there are significant challenges to overcome.

First, deep learning algorithms require a large number of examples in order to learn complex patterns and make accurate predictions, which can be a challenge in the field of genomics where data, especially from validated enhancers, is limited. Because the sequence syntax and logic within enhancers are complex and context dependent, understanding the regulatory code that determines when and where genes are expressed in animals requires access to a large amount of data in diverse cell types and time points. There is a paucity of large datasets and enhancers with validated activity in humans tend to be restricted to a handful of cell lines (i.e., K562) and a subset of evolutionarily conserved enhancers between human and mouse [211].

The use of large-scale datasets, including those generated through consortium initiatives like ENCODE and the NIH Roadmap Epigenomics Consortium, will continue to be a valuable resource for machine learning approaches. Developments in high-throughput molecular validation methods to allow for more cell types to be tested will improve data availability for machine learning models [170, 175, 176].

Second, understanding the specific biological features that drive model predictions and the decision-making process is an area of active research [215]. The development of more accurate and interpretable machine learning approaches can lead to a greater understanding of the complexities of enhancer function and the identification of new regulatory elements and mechanisms. NLP models may generally improve interpretability. Another exciting area of research is latent text-to-image generative models that are being applied to design novel cell-type specific regulatory elements, which when combined with molecular validation can help further elucidate cell-type specific regulatory codes [227].

Proteomics can be used to validate promising findings to gain novel biological insights. The integration of machine learning with experimental validation will be key

to fully realizing the potential of these approaches to decipher the intricate relationship between enhancer sequences and activity.

Third, while deep learning algorithms can be highly accurate, they do not always generalize well to new datasets. To be able to accurately transfer knowledge across cell types in different species would be a valuable tool. The use of transfer learning, which involves pre-training a model on a large dataset and then fine-tuning it on a smaller dataset may improve the performance of deep learning models for predicting TF binding sites [209].

Finally, while the development of predictive models that can identify and predict the activity of endogenous and synthetic *cis*-regulatory elements provides an important framework for understanding enhancers, a unified definition of context-specific enhancer activity based on interpretable sequence rules would serve as a basic organizational principle of the regulatory genome.

Conclusions

A major goal in genetics is to elucidate enhancer sequences to better understand how the genome encodes cell and organismal traits. Enhancers are characterized by features that make them highly flexible and evolvable, including redundancy, modularity, sequence degeneracy, and binding suboptimality. These features provide robustness, but they also make it challenging to decipher the underlying principles of enhancer function.

In silico methods, such as machine learning, combined with single-cell approaches offer new avenues to study enhancers and understand the relationship between their sequence and activity in different in vivo contexts, including rare and transient cell states. While these methods have been successful in identifying candidate enhancers and their gene networks, we are still in the early stages of developing biologically meaningful sequence models that can accurately predict enhancer activity in specific cell types and at specific time points.

Ongoing developments in technology and data collection, including in areas such as single cell genomics, will be critical for advancing our understanding of enhancers and other *cis*-regulatory elements. By leveraging these advances, we can build predictive and interpretable frameworks for understanding the sequence basis of enhancers to gain insights into their role in shaping organismal phenotypes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02955-4.

Additional file 1. Review history.

Acknowledgements
We thank Adam Siepel, Lithin Louis, Liam Reynolds, and Jack Clarke for their valuable feedback.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 1.

Authors' contributions
GDS, WHC, and ESW drafted the initial manuscript. ESW conceptualized, coordinated, and extensively revised the manuscript. All authors contributed to visualizations. The author(s) read and approved the final manuscript.

Funding
GDS is supported by an Australian Government Research Training Program Scholarship. PCP is supported by a UNSW International Postgraduate Award. ESW is supported by a National Health and Medical Research Council Investigator Grant (GNT2009309), Australian Research Council Discovery Project (DP200100250), and a Snow Medical Fellowship.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 18 July 2022 Accepted: 28 April 2023
Published online: 12 May 2023

References

1. Wray GA. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 2003;20:1377–419.
2. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.
3. Wray GA. Transcriptional regulation and the evolution of development. Int J Dev Biol. 2003;47:675–84.
4. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981;27:299–308.
5. Long HK, Prescott SL, Wysocka J. Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell. 2016;167:1170–87.
6. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:206–16.
7. Noonan JP, McCallion AS. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet. 2010;11:1–23.
8. Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, et al. Coding exons function as tissue-specific enhancers of nearby genes. Genome Res. 2012;22:1059–68.
9. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488:116–20.
10. Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, et al. Genome-wide map of regulatory interactions in the human genome. Genome Res. 2014;24:1905–17.
11. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.
12. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
13. Li X-Y, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol. 2011;12:R34.
14. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
15. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515:355–64.
16. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
17. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.
18. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54:940–9.
19. Patel ZM, Hughes TR. Global properties of regulatory sequences are predicted by transcription factor recognition mechanisms. Genome Biol. 2021;22:285.
20. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9.
21. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011;21:2167–80.
22. Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics. 2017;33:1930–6.
23. Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, et al. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. Genome Res. 2012;22:2290–301.
24. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet. 2015;47:955–61.
25. Atak ZK, Taskiran I, Demeulemeester J, Flerin C, Mauduit D, Minnoye L, et al. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. Genome Res. 2021;31:1082–96.
26. Minnoye L, Taskiran I, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, et al. Cross-species analysis of enhancer logic using deep learning. Genome Res. 2020;30:1815–34.
27. Sahu B, Hartonen T, Pihlajamaa P, Wei B, Dave K, Zhu F, et al. Sequence determinants of human gene regulatory elements. Nat Genet. 2022;54:283–94.

28. Galupa R, Alvarez-Canales G, Borst NO, Fuqua T, Gandara L, Misunou N, et al. Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development. Dev Cell. 2023;58:51-62.e4.
29. Weber B, Zicola J, Oka R, Stam M. Plant enhancers: a call for discovery. Trends Plant Sci. 2016;21:974–87.
30. Sun J, He N, Niu L, Huang Y, Shen W, Zhang Y, et al. Global quantitative mapping of enhancers in rice by STARR-seq. Genomics Proteomics Bioinformatics. 2019;17:140–53.
31. Sijacic P, Bajic M, McKinney EC, Meagher RB, Deal RB. Chromatin accessibility changes between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. Plant J Cell Mol Biol. 2018;94:215–31.
32. Zhang W, Wu Y, Schnable JC, Zeng Z, Freeling M, Crawford GE, et al. High-resolution mapping of open chromatin in the rice genome. Genome Res. 2012;22:151–62.
33. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. Plant Cell. 2022;34:718–41.
34. Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. Trends Mol Med. 2021;27:1060–73.
35. Kvon EZ, Zhu Y, Kelman G, Novak CS, Plajzer-Frick I, Kato M, et al. Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants. Cell. 2020;180:1262-1271.e15.
36. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet. 2019;20:437–55.
37. Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev Cell. 2009;16:47–57.
38. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012;149:1233–44.
39. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014;512:96–100.
40. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, et al. Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet. 2011;7:e1001343.
41. Tippens ND, Vihervaara A, Lis JT. Enhancer transcription: what, where, when, and why? Genes Dev. 2018;32:1–3.
42. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol. 2020;27:521–8.
43. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet. 2016;17:207–23.
44. Tome JM, Tippens ND, Lis JT. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nat Genet. 2018;50:1533–41.
45. Tsai PF, Dell'Orso S, Rodriguez J, Vivanco KO, Ko KD, Jiang K, et al. A muscle-specific enhancer RNA mediates cohesin recruitment and regulates transcription in trans. Mol Cell. 2018;71:129-141.e8.
46. Hsieh CL, Fei T, Chen Y, Li T, Gao Y, Wang X, et al. Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. Proc Natl Acad Sci U S A. 2014;111:7319–24.
47. Mousavi K, Zare H, Dell'orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, et al. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. Mol Cell. 2013;51:606–17.
48. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48:488–96.
49. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A phase separation model predicts key features of transcriptional control. Cell. 2017;169:13–23.
50. Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. Science. 2018;361. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29930091.
51. Karr JP, Ferrie JJ, Tjian R, Darzacq X. The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer-promoter communication. Genes Dev. 2022;36:7–16.
52. Benabdallah NS, Williamson I, Illingworth RS, Kane L, Boyle S, Sengupta D, et al. Decreased enhancer-promoter proximity accompanying enhancer activation. Mol Cell. 2019;76:473-484.e7.
53. Alexander JM, Guan J, Li B, Maliskova L, Song M, Shen Y, et al. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. eLife. 2019;8:e41769.
54. Gorbovytska V, Kim SK, Kuybu F, Gotze M, Um D, Kang K, et al. Enhancer RNAs stimulate Pol II pause release by harnessing multivalent interactions to NELF. Nat Commun. 2022;13:2429.
55. Blick AJ, Mayer-Hirshfeld I, Malibiran BR, Cooper MA, Martino PA, Johnson JE, et al. The capacity to act in trans varies among Drosophila enhancers. Genetics. 2016;203:203–18.
56. Geyer PK, Green MM, Corces VG. Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila. EMBO J. 1990;9:2247–56.
57. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107:21931–6.
58. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457:854–8.
59. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.
60. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020;584:244–51.
61. Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods. 2015;12:433–8.
62. Tippens ND, Liang J, Leung AK-Y, Wierbowski SD, Ozer A, Booth JG, et al. Transcription imparts architecture, function and logic to enhancer units. Nat Genet. 2020;52:1067–75.

63. Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends Genet. 2015;31:426–33.
64. Wang Z, Chu T, Choate LA, Danko CG. Identification of regulatory elements from nascent transcription using dREG. Genome Res. 2019;29:293–303.
65. Kouno T, Moody J, Kwon AT-J, Shibayama Y, Kato S, Huang Y, et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nat Commun. 2019;10:360.
66. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46:1311–20.
67. Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, et al. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. 2018;32:42–57.
68. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biol. 2021;22:108.
69. Cajigas I, Chakraborty A, Swyter KR, Luo H, Bastidas M, Nigro M, et al. The Evf2 ultraconserved enhancer lncRNA functionally and spatially organizes megabase distant genes in the developing forebrain. Mol Cell. 2018;71:956-972.e9.
70. Lee J-H, Wang R, Xiong F, Krakowiak J, Liao Z, Nguyen PT, et al. Enhancer RNA m6A methylation facilitates transcriptional condensate formation and gene activation. Mol Cell. 2021;81:3368-3385.e9.
71. Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. BMC Genomics. 2019;20:511.
72. Hornblad A, Bastide S, Langenfeld K, Langa F, Spitz F. Dissection of the Fgf8 regulatory landscape by in vivo CRISPR-editing reveals extensive intra- and inter-enhancer redundancy. Nat Commun. 2021;12:439.
73. Bhatia S, Jan Kleinjan D, Uttley K, Mann A, Dellepiane N, Bickmore WA. Quantitative spatial and temporal assessment of regulatory element activity in zebrafish. eLife. 2021;10:e65601.
74. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell. 2006;124:47–59.
75. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, 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;99:757–62.
76. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, et al. A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell. 2012;148:473–86.
77. Uhl JD, Zandvakili A, Gebelein B. A Hox Transcription factor collective binds a highly conserved distal-less cis-regulatory module to generate robust transcriptional outcomes. PLoS Genet. 2016;12:e1005981.
78. Jindal GA, Farley EK. Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev Cell. 2021;56:575–87.
79. Kulkarni MM, Arnosti DN. Information display by transcriptional enhancers. Development. 2003;130:6569–75.
80. Thanos D, Maniatis T. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. Cell. 1995;83:1091–100.
81. King DM, Hong CKY, Shepherdson JL, Granas DM, Maricque BB, Cohen BA. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife. 2020;9:e41279.
82. Ng FS, Schutte J, Ruau D, Diamanti E, Hannah R, Kinston SJ, et al. Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. Nucleic Acids Res. 2014;42:13513–24.
83. Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc Natl Acad Sci U S A. 2016;113:6508–13.
84. Singh G, Mullany S, Moorthy SD, Zhang R, Mehdi T, Tian R, et al. A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. Genome Res. 2021;31:564–75.
85. Wong ES, Zheng D, Tan SZ, Bower NI, Garside V, Vanwalleghem G, et al. Deep conservation of the enhancer regulatory code in animals. Science. 2020;370:eaax8137.
86. Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet. 2013;45:1021–8.
87. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. Perrimon N, editor. PLoS Genet. 2008;4:e1000106.
88. Schnepf M, von Reutern M, Ludwig C, Jung C, Gaul U. Transcription factor binding affinities and DNA shape readout. iScience. 2020;23:101694.
89. Samee MdAH, Bruneau BG, Pollard KS. A De novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. Cell Syst. 2019;8:27-42.e6.
90. Sielemann J, Wulf D, Schmidt R, Brautigam A. Local DNA shape is a general principle of transcription factor binding specificity in Arabidopsis thaliana. Nat Commun. 2021;12:6549.
91. Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. 2013;3:1093–104.
92. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28473536.
93. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160:554–66.
94. Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, et al. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat Genet. 2014;46:685–92.
95. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010;42:806–10.

96. Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell. 2013;155:1521–31.
97. Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. Many human accelerated regions are developmental enhancers. Philos Trans R Soc B Biol Sci. 2013;368:20130025.
98. Moon JM, Capra JA, Abbot P, Rokas A. Signatures of recent positive selection in enhancers across 41 human tissues. G3. 2019;9:2761–74.
99. Snetkova V, Pennacchio LA, Visel A, Dickel DE. Perfect and imperfect views of ultraconserved sequences. Nat Rev Genet. 2022;23:182–94.
100. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science. 2004;304:1321–5.
101. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet. 2008;40:158–60.
102. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans — mechanisms and functional implications. Nat Rev Genet. 2014;15:221–33.
103. Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond AM, et al. LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. Genome Biol. 2021;22:62.
104. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A. 2007;104:8005–10.
105. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008;18:1752–62.
106. White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci. 2013;110:11952–7.
107. Duque T, Sinha S. What does it take to evolve an enhancer? A simulation-based study of factors influencing the emergence of combinatorial regulation. Genome Biol Evol. 2015;7:1415–31.
108. Smith RP, Riesenfeld SJ, Holloway AK, Li Q, Murphy KK, Feliciano NM, et al. A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. Genome Biol. 2013;14:R72.
109. Zemojtel T, Kielbasa SM, Arndt PF, Behrens S, Bourque G, Vingron M. CpG deamination creates transcription factor-binding sites with high efficiency. Genome Biol Evol. 2011;3:1304–11.
110. Stone JR, Wray GA. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol. 2001;18:1764–70.
111. Schwaiger M, Schönauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. Genome Res. 2014;24:639–50.
112. Royo JL, Maeso I, Irimia M, Gao F, Peter IS, Lopes CS, et al. Transphyletic conservation of developmental regulatory state in animal evolution. Proc Natl Acad Sci U S A. 2011;108:14186–91.
113. Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, et al. Genome-wide identification of conserved regulatory function in diverged sequences. Genome Res. 2011;21:1139–49.
114. Ahituv N, Prabhakar S, Poulin F, Rubin EM, Couronne O. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum Mol Genet. 2005;14:3057–63.
115. Irimia M, Tena JJ, Alexis MS, Fernandez-Miñan A, Maeso I, Bogdanovic O, et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. 2012;22:2356–67.
116. Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, et al. Transcriptional features of genomic regulatory blocks. Genome Biol. 2009;10:R38.
117. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 2007;17:545–55.
118. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat Commun. 2017;8:1–13.
119. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science. 2006;312:276–9.
120. Sebe-Pedros A, Ballare C, Parra-Acero H, Chiva C, Tena JJ, Sabido E, et al. The dynamic regulatory genome of Capsaspora and the origin of animal multicellularity. Cell. 2016;165:1224–37.
121. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.
122. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature. 2010;466:720–6.
123. Cornejo-Paramo P, Roper K, Degnan SM, Degnan BM, Wong ES. Distal regulation, silencers, and a shared combinatorial syntax are hallmarks of animal embryogenesis. Genome Res. 2022;32:474–87.
124. McCole RB, Erceg J, Saylor W, Wu CT. Ultraconserved elements occupy specific arenas of three-dimensional mammalian genome organization. Cell Rep. 2018;24:479–88.
125. Viturawong T, Meissner F, Butter F, Mann M. A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation. Cell Rep. 2013;5:531–45.
126. Glassford WJ, Rebeiz M. Assessing constraints on the path of regulatory sequence evolution. Philos Trans R Soc Lond B Biol Sci. 2013;368:20130026.
127. Snetkova V, Ypsilanti AR, Akiyama JA, Mannion BJ, Plajzer-Frick I, Novak CS, et al. Ultraconserved enhancer function does not require perfect sequence conservation. Nat Genet. 2021;53:521–8.
128. Liu J, Viales RR, Khoueiry P, Reddington JP, Girardot C, Furlong EEM, et al. The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection. Genome Res. 2021;31:1573–81.

129. Cornejo-Páramo P, Petrova V, Zhang X, Young RS, Wong ES. Enhancer turnover in cancer and species evolution are associated with DNA replication timing. bioRxiv; 2022. Available from: https://www.biorxiv.org/content/10.1101/2022.12.22.521323v1.
130. Kaplow IM, Lawler AJ, Schäffer DE, Srinivasan C, Sestili HH, Wirthlin ME, et al. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. Science. 2023;380:eabm7993.
131. Kaplow IM, Schäffer DE, Wirthlin ME, Lawler AJ, Brown AR, Kleyman M, et al. Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. BMC Genomics. 2022;23:291.
132. Kitano H. Biological robustness. Nat Rev Genet. 2004;5:826–37.
133. Spivakov M. Spurious transcription factor binding: non-functional or genetically redundant? BioEssays. 2014;36:798–806.
134. Li S, Kvon EZ, Visel A, Pennacchio LA, Ovcharenko I. Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation. Genome Biol. 2019;20:140.
135. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012;30:265–70.
136. Hare EE, Peterson BK, Eisen MB. A careful look at binding site reorganization in the even-skipped enhancers of Drosophila and sepsids. PLoS Genet. 2008;4:e1000268.
137. Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, Wang S, et al. Low affinity binding site clusters confer Hox specificity and regulatory robustness. Cell. 2015;160:191–203.
138. Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. Suboptimization of developmental enhancers. Science. 2015;350:325–8.
139. Ramos AI, Barolo S. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos Trans R Soc Lond B Biol Sci. 2013;368:20130018.
140. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annu Rev Cell Dev Biol. 2019;35:357–79.
141. Le DD, Shimko TC, Aditham AK, Keys AM, Longwell SA, Orenstein Y, et al. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proc Natl Acad Sci U S A. 2018;115:E3702–11.
142. Hong JW, Hendrix DA, Levine MS. Shadow enhancers as a source of evolutionary novelty. Science. 2008;321:1314.
143. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow enhancers foster robustness of Drosophila gastrulation. Curr Biol. 2010;20:1562–7.
144. Kvon EZ, Waymack R, Gad M, Wunderlich Z. Enhancer redundancy in development and disease. Nat Rev Genet. 2021;22:324–36.
145. Floc'hlay S, Wong ES, Zhao B, Viales RR, Thomas-Chollier M, Thieffry D, et al. Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. Genome Res. 2021;31:211–24.
146. Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature. 2018;554:239–43.
147. Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature. 2010;466:490–3.
148. Waymack R, Fletcher A, Enciso G, Wunderlich Z. Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic. eLife. 2020;9:e59351.
149. Tsai A, Alves MR, Crocker J. Multi-enhancer transcriptional hubs confer phenotypic robustness. Arnosti DN, Tyler JK, DePace AH, Garcia H, editors. eLife. 2019;8:e45325.
150. Cannavo E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, et al. Shadow enhancers are pervasive features of developmental regulatory networks. Curr Biol. 2016;26:38–51.
151. Preger-Ben Noon E, Davis FP, Stern DL. Evolved repression overcomes enhancer robustness. Dev Cell. 2016;39:572–84.
152. Ibarra IL, Hollmann NM, Klaus B, Augsten S, Velten B, Hennig J, et al. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. Nat Commun. 2020;11:124.
153. Guo Y, Gifford DK. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. BMC Genomics. 2017;18:45.
154. Tanay A. Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res. 2006;16:962–72.
155. Wong ES, Thybert D, Schmitt BM, Stefflova K, Odom DT, Flicek P. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. Genome Res. 2015;25:167–78.
156. Payne JL, Wagner A. The robustness and evolvability of transcription factor binding sites. Science. 2014;343:875–7.
157. Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. Nat Commun. 2018;9:1530.
158. Ohno S. Evolution by gene duplication. Berlin: Springer-Verlag; 1970.
159. Wong ES, Schmitt BM, Kazachenka A, Thybert D, Redmond A, Connor F, et al. Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nat Commun. 2017;8:1092.
160. Ryan GE, Farley EK. Functional genomic approaches to elucidate the role of enhancers during development. WIREs Syst Biol Med. 2020;12:e1467.
161. Kinney JB, McCandlish DM. Massively parallel assays and quantitative sequence-function relationships. Annu Rev Genomics Hum Genet. 2019;20:99–127.
162. Kreimer A, Ashuach T, Inoue F, Khodaverdian A, Deng C, Yosef N, et al. Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. Nat Commun. 2022;13:1504.
163. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat Biotechnol. 2020;38:56–65.
164. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30:271–7.
165. Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics. 2015;106:159–64.
166. Hong CKY, Cohen BA. Genomic environments scale the activities of diverse core promoters. Genome Res. 2022;32:85–96.

167. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell. 2013;154:914–27.
168. Maricque BB, Dougherty JD, Cohen BA. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. Nucleic Acids Res. 2017;45:e16.
169. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat Methods. 2020;17:1083–91.
170. Akerberg BN, Gu F, VanDusen NJ, Zhang X, Dong R, Li K, et al. A reference map of murine cardiac transcription factor chromatin occupancy identifies dynamic and conserved enhancers. Nat Commun. 2019;10:4907.
171. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019;10:3583.
172. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell. 2016;165:1530–45.
173. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013;23:800–11.
174. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A. 2012;109:19498–503.
175. Lalanne J-B, Regalado SG, Domcke S, Calderon D, Martin B, Li T, et al. Multiplex profiling of developmental enhancers with quantitative, single-cell expression reporters. 2022. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.12.10.519236. Cited 2022 Dec 12.
176. Zhao S, Hong CKY, Myers CA, Granas DM, White MA, Corbo JC, et al. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. Nat Genet. 2023;55:346–54. Nature Publishing Group.
177. Li K, Liu Y, Cao H, Zhang Y, Gu Z, Liu X, et al. Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. Nat Commun. 2020;11:485.
178. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science. 2016;354:769–73.
179. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014;159:647–61.
180. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet. 2017;49:1602–12.
181. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152:1173–83.
182. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33:510–7.
183. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020;21:292–310.
184. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176:377-390.e19.
185. Schraivogel D, Gschwind AR, Milbank JH, Leonce DR, Jakob P, Mathur L, et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat Methods. 2020;17:629–35.
186. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.
187. Thakore PI, Kwon JB, Nelson CE, Rouse DC, Gemberling MP, Oliver ML, et al. RNA-guided transcriptional silencing in vivo with S. aureus CRISPR-Cas9 repressors. Nat Commun. 2018;9:1674.
188. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51:12–8.
189. Cochran K, Srivastava D, Shrikumar A, Balsubramani A, Hardison RC, Kundaje A, et al. Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. Genome Res. 2022;32:512–23.
190. Chen L, Fish AE, Capra JA. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLoS Comput Biol. 2018;14:e1006484.
191. Minnoye L, Taskiran II, Mauduit D, Fazio M, Aerschot LV, Hulselmans G, et al. Cross-species analysis of enhancer logic using deep learning. Genome Res. 2020. https://doi.org/10.1101/gr.260844.120.
192. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12:931–4.
193. de Almeida BP, Reiter F, Pagani M, Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet. 2022;54:613–24.
194. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
195. Bergman DT, Jones TR, Liu V, Ray J, Jagoda E, Siraj L, et al. Compatibility rules of human enhancer and promoter sequences. Nature. 2022;607:176–84.
196. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol. 2014;10:e1003711.
197. Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, et al. Integrating diverse datasets improves developmental enhancer prediction. PLoS Comput Biol. 2014;10:e1003677.
198. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol. 2013;9:e1002968.
199. Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics. 2017;18:478.
200. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
201. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44:e107.

202. Zheng A, Lamkin M, Zhao H, Wu C, Su H, Gymrek M. Deep neural networks identify sequence context features predictive of transcription factor binding. Nat Mach Intell. 2021;3:172–80.
203. Chen L, Capra JA. Learning and interpreting the gene regulatory grammar in a deep learning framework. PLoS Comput Biol. 2020;16:e1008334.
204. Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics. 2010;26:1579–86.
205. Bu H, Gan Y, Wang Y, Zhou S, Guan J. A new method for enhancer prediction based on deep belief network. BMC Bioinformatics. 2017;18:418.
206. Avsec Z, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet. 2021;53:354–66.
207. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.
208. Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell. 2021;184:5053-5069.e23.
209. Novakovsky G, Saraswat M, Fornes O, Mostafavi S, Wasserman WW. Biologically relevant transfer learning improves transcription factor binding prediction. Genome Biol. 2021;22:280.
210. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. 2019;47:D729–35.
211. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35:D88-92.
212. Shrikumar A, Prakash E, Kundaje A. GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. Bioinformatics. 2019;35:i173–82.
213. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.
214. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232. Institute of Mathematical Statistics.
215. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2022;24:125–37.
216. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. Available from: http://arxiv.org/abs/1705.07874.
217. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. 2020. Available from: http://arxiv.org/abs/1811.00416.
218. Le NQK, Ho Q-T, Nguyen V-N, Chang J-S. BERT-Promoter: an improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. Comput Biol Chem. 2022;99:107732.
219. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389–403.
220. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. Available from: http://arxiv.org/abs/1810.04805.
221. Karollus A, Mauermeier T, Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biol. 2023;24:56.
222. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinform. 2021;22:177.
223. Koo PK, Eddy SR. Representation learning of genomic sequence motifs with convolutional neural networks. PLoS Comput Biol. 2019;15:e1007560.
224. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Doina P, Yee Whye T, editors. Proceedings of Machine Learning Research: PMLR. 2017. p. 3145–3153. Available from: https://proceedings.mlr.press/v70/shrikumar17a.html.
225. Novakovsky G, Fornes O, Saraswat M, Mostafavi S, Wasserman WW. ExplaiNN: interpretable and transparent neural networks for genomics. bioRxiv; 2022. Available from: https://www.biorxiv.org/content/10.1101/2022.05.20.492818v2.
226. Ameen M, Sundaram L, Shen M, Banerjee A, Kundu S, Nair S, et al. Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. Cell. 2022;185:4937-4953.e23.
227. Taskiran II, Spanier KI, Christiaens V, Mauduit D, Aerts S. Cell type directed design of synthetic enhancers. bioRxiv; 2022. p. 2022.07.26.501466. Available from: https://www.biorxiv.org/content/10.1101/2022.07.26.501466v1.

Publisher's Note