Computational and Structural Biotechnology Journal 20 (2022) 2728–2744

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Analysis of the landscape of human enhancer sequences in biological databases

Juan Mulero Hernández, Jesualdo Tomás Fernández-Breis *

Dept. Informática y Sistemas, Universidad de Murcia, CEIR Campus Mare Nostrum, IMIB-Arrixaca, Spain

ARTICLE INFO

Article history:
Received 28 March 2022
Received in revised form 20 May 2022
Accepted 21 May 2022
Available online 30 May 2022

Keywords:
Bioinformatics
Gene regulation
Enhancers
Human
Biological databases

ABSTRACT

The process of gene regulation extends as a network in which both genetic sequences and proteins are involved. The levels of regulation and the mechanisms involved are multiple. Transcription is the main control mechanism for most genes, being the downstream steps responsible for refining the transcription patterns. In turn, gene transcription is mainly controlled by regulatory events that occur at promoters and enhancers. Several studies are focused on analyzing the contribution of enhancers in the development of diseases and their possible use as therapeutic targets. The study of regulatory elements has advanced rapidly in recent years with the development and use of next generation sequencing techniques. All this information has generated a large volume of information that has been transferred to a growing number of public repositories that store this information. In this article, we analyze the content of those public repositories that contain information about human enhancers with the aim of detecting whether the knowledge generated by scientific research is contained in those databases in a way that could be computationally exploited. The analysis will be based on three main aspects identified in the literature: types of enhancers, type of evidence about the enhancers, and methods for detecting enhancer-promoter interactions. Our results show that no single database facilitates the optimal exploitation of enhancer data, most types of enhancers are not represented in the databases and there is need for a standardized model for enhancers. We have identified major gaps and challenges for the computational exploitation of enhancer data.

1. Introduction

Enhancers are distal cis-regulatory sequences capable of increasing the transcription of genes that they regulate independently of their orientation and distance to the transcription start site (TSS) [1–3]. Moreover, they have been shown to be fundamental sequences in the regulation of genes and processes of relevance such as cell identity and disease development. In fact, the term enhancerophaties has been used to refer to diseases associated with these sequences, and they have been studied as possible therapeutic targets [4–8].

Literature estimates that the human genome contains more enhancers than protein-coding genes, because one gene can be regulated by multiple enhancers, although one enhancer can also control multiple genes [3,9,10] (Fig. 1). Some studies have found that each enhancer interacts with approximately two promoters and each promoter interacts with 4–5 enhancer elements [9]. Furthermore, the specificity of species and tissue implies that different combinations of enhancers can be used to control the expression of a gene and that they can be different depending on the cellular state and environmental factors [3,9,11,12]. Even at the enhancer level, the existence of multiple transcription factor binding sequences (TFBS) means that combinations of transcription factors (TF) and mechanisms of action can be even more varied and context dependent [13].

Initially, enhancer sequences have followed a functional definition. They have traditionally been characterised as nucleosome-free regions (NFR or DHS) enriched with TFBS that allow the recruitment of molecular elements. In turn, they collaborate in the gene transcription process when these molecular elements can interact through chromatin loops that approximate the sequences or through the formation of hubs [14–16] (Fig. 1). Multiple mechanisms have been observed and suggested to explain how enhancers are able to increase the expression of target genes [17]. We find modes of action based on the nature of the sequence that performs the enhancer action (enhancer action based on the DNA sequence or transcribed eRNA), the environment of action

* Corresponding author.
E-mail address: jfernand@um.es (J.T. Fernández-Breis).

https://doi.org/10.1016/j.csbj.2022.05.045
2001-0370/

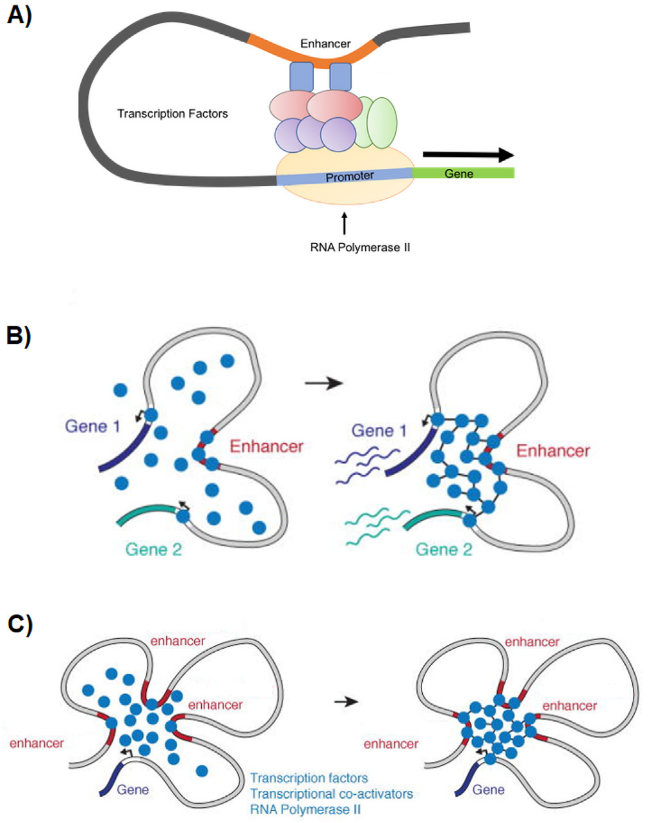

Fig. 1. Traditional chromatin loop model (A). The enhancer physically interacts with the promoter through chromatin flexibility and mechanisms like the loop extrusion model. In this way, the enhancer can provide molecular elements that increase transcription of the target gene. Alternatively, in the hub model, the spatial proximity of multiple regulatory sequences allows the recruitment of high concentrations of molecular elements that can generate a network and a microenvironment, even phase-separated, that increases the transcription of target genes. This model can explain phenomena like the regulation of multiple genes by the same sequence (B) and the regulation of the same gene by multiple sequences (C). Modified from www.addgene.org and [42].

(cis/trans action), the mode of linking between regulatory sequences (chromatin loop, transcription factories or hubs, facilitate tracking, linking/chaining), the mechanism that initiates the enhancer action (recruitment of transcription factors, cofactors, chromatin modifiers, RNAPII transfer, liberation of the transcriptional pause and eRNA action) or the consequence of the enhancer activity (increase the initiation or elongation of the transcription or reduce the pause of the transcription). It has also been proposed that they can increase splicing, polyadenylation, transcription termination rate and RNAPII recycling, but the implication of enhancers in post-transcriptional regulatory steps requires more study [11,17–23]. In short, enhancers have the capacity to regulate by increasing the transcription of a gene, increasing the activity of a promoter or providing essential information that the promoter does not provide.

With the development of genomics and next generation sequencing (NGS), studies have tried to identify genomic features in these sequences to study their function, their mechanisms of action and to perform a massive screening of these sequences in the genome using these correlated physical properties [24,25]. In this task, high-throughput genomics methods have allowed us to study in depth the chromatin properties of enhancers [26] and the transcripts that they can generate (eRNA) [27–30], while chromatin conformation capture and high-resolution microscopy techniques have allowed the determination of distal chromosomal contacts between promoter and enhancer sequences (EPI), as well as the mechanisms of action [31–34]. However, deeper exploration of these properties has also shown that there is no homogeneous profile of features in the enhancers [26,35] and, therefore, the different methodologies provide a partial view of the regulatory landscape [36]. The same situation applies to the identification of EPI [37]. In addition, the different characteristics identified in the enhancers have also generated different classifications and terminologies that have extended in the literature, whose validity of use generates controversy due to the lack of consensus in the definition of the enhancers [12] and the absence of a controlled vocabulary by models of knowledge representation.

The study of regulatory elements has advanced quickly in recent years with the development and use of new techniques, and this progress has been paralleled with the clinical purpose of detecting and studying diseases [25,38–41].

As a result, a large volume of data and knowledge about human enhancer sequences has been generated and has been stored in a growing number of biological databases. Given the increasing bioinformatics processing of these data, reviewing the content of the databases will show how well the content in databases covers the knowledge generated by the scientific community and which are the main limitations for the computational exploitation of enhancer data. Hence, we will first describe a model for representing human enhancer sequences derived from the analysis of the literature. That model will drive the analysis of the content of 25 biological databases, which will focus on three main aspects: types of enhancers, type of evidence about the enhancers, and methods for detecting enhancer-promoter interactions. Finally, we identify the main challenges in this area.

This work aims at contributing to describe the current landscape of human enhancers data and to the development of the Gene Regulation Knowledge Commons targeted by the GREEKC COST Action (https://greekc.org/). The GREEKC focus is on the generation, curation and analysis of data and knowledge about gene regulation processes. The current article is focused on the data resources, but discussion on the need for enhancer data interoperability is also provided in this article.

2. Materials and methods

2.1. A model for representing enhancers

After a review of the literature, we propose a model to represent the fundamental information about enhancers (Fig. 2). In the following subsections we describe the main elements which should be covered and which will drive our analysis of bioinformatics data resources: types of enhancer, methodologies to generate evidence, enhancer-promoter interactions (EPI) and other annotations of interest. Furthermore, any enhancer must have, at least, the coordinates that allow mapping this sequence in the genome and the biosample. As we will see below, the type of enhancer may vary according to the biological sample, because enhancers are regulatory sequences specific to cell type, but also to environmental stimuli [3,13].

2.1.1. Types of enhancers

Enhancers do not have homogeneous characteristics, but have a variable profile [26]. Enhancers are dynamic and specific of species and tissue, even TFBS that make up the enhancers can be specific (pleiotropic sequences) [13]. Therefore, annotating a sequence as enhancer is not sufficient to represent the description of its biology and multiple typologies have emerged in the scientific literature to classify different patterns of features. Fig. 3 includes the types of enhancers which have been described in the literature and shows

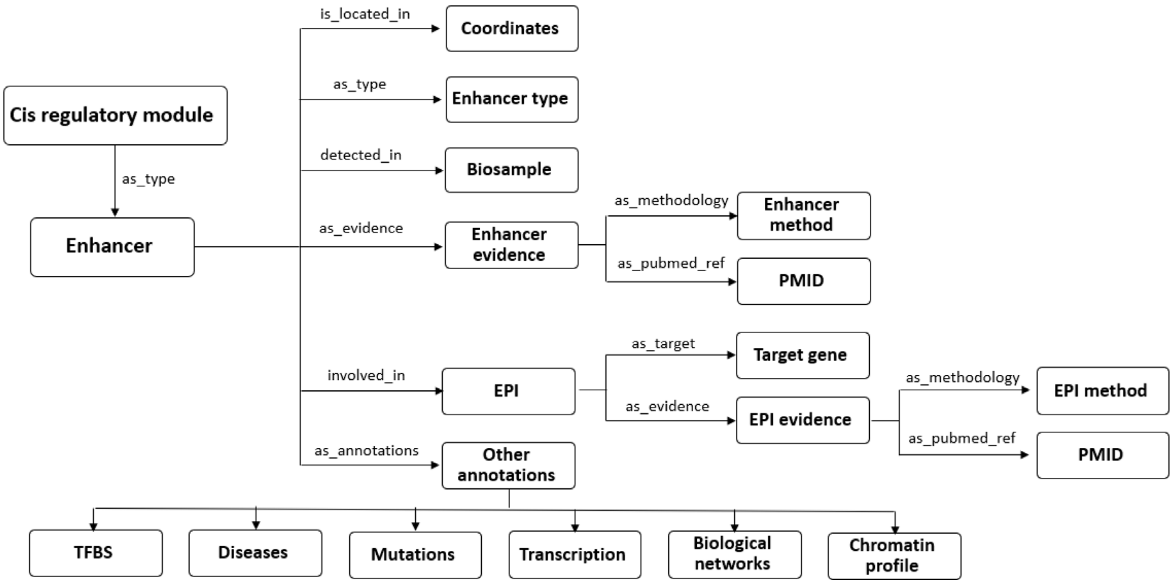

Fig. 2. Proposed model for the representation of enhancers. Each enhancer is located in a region of the genome and belongs to one or more classifications of enhancers, which may differ according to the biological sample in question, because enhancers are sequence specific. The identification of the enhancer has evidence derived from the methodology used and must have a bibliographic reference that allows to verify the information. As regulatory sequences, enhancers regulate genes and this regulation is also supported by evidence. In addition, the enhancers can be enriched with information of interest such as their link to diseases or the TFBS that compose the sequence.

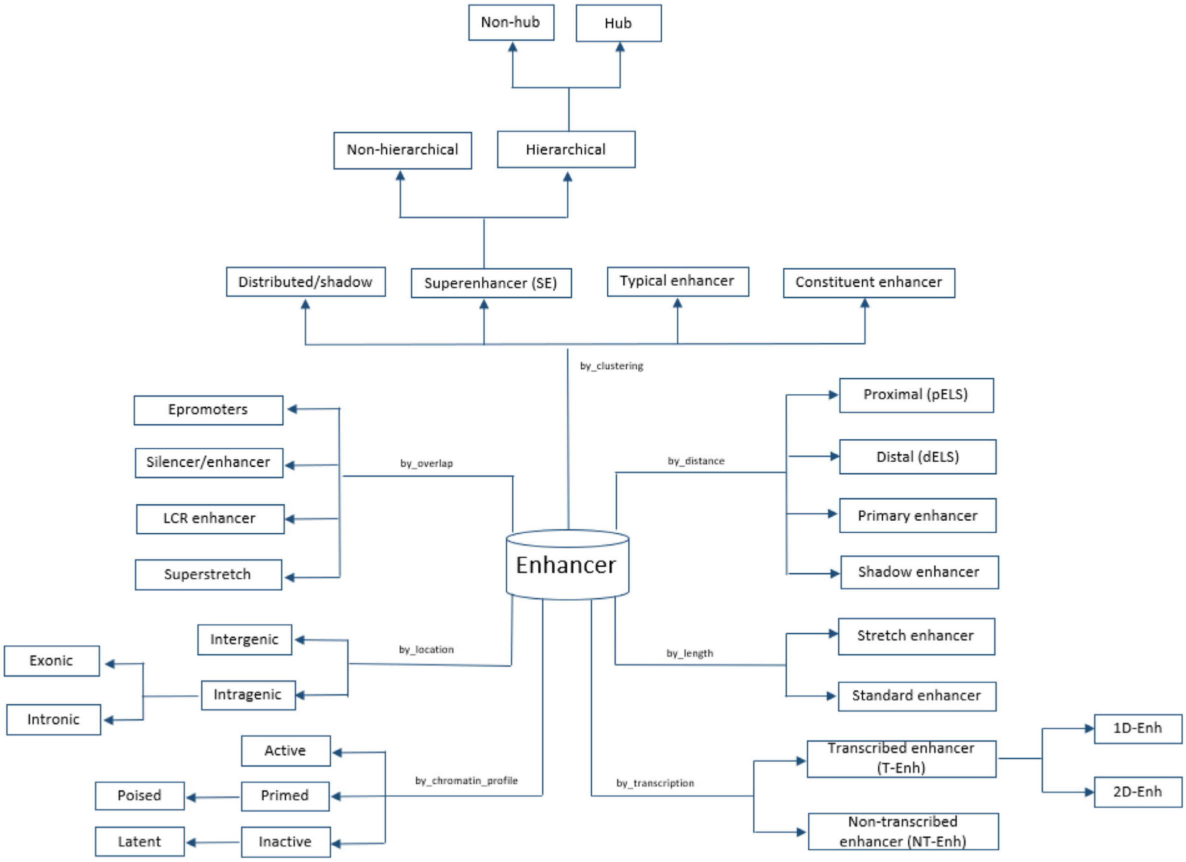

Fig. 3. Classification of the main types of enhancers found in the literature. The characteristics of enhancers do not have a homogeneous profile. For this reason, we find different classifications in the literature, which have been compiled in this figure. Each classification is based on different properties, so an enhancer can belong to several types at the same time.

different ways of classifying and subtyping enhancers, so one enhancer can belong to different types at the same time. We describe next those types:

Enhancers by distance: This classification is based on the distance of the enhancer to promoters or target genes. **Primary and shadow enhancers** are likely to be one of the most frequently mentioned types of enhancer in the related literature [11,6]. Initially, enhancers that were close to promoters were considered as primary or principal enhancers in the regulation of gene expression, while more remote enhancers that exhibited the same regulatory activity were considered as secondary or redundant [43]. However, studies have shown that the genomic characteristics of primary and shadow enhancers do not differ significantly [11]. The selective restriction of shadow enhancers can be as great or even greater than that of other enhancers [44] and they can act simultaneously with other enhancers and over large distances [45]. Other studies have also shown that they can contribute to the accuracy and robustness of gene transcription against environmental and genetic variability, as well as stochastic perturbations, by reducing the level of transcription noise [22,46,44,47]. Therefore, their contribution may be relevant, meaning that this classification based on the position of the sequence with respect to the target gene is not the most useful. In addition, because one enhancer can regulate more than one gene, an enhancer could be labeled as both a primary and a shadow enhancer depending on the gene.

Proximal and distal enhancers are part of this hierarchy, and they differ on the distance to the target gene. If the distance is larger than a certain value, then the enhancer is classified as distal, otherwise it is proximal. There is no consensus on the threshold distance. For example, the threshold in SCREEN ENCODE is 2 kb.

Enhancers by location: This classification is based on the location of the enhancers in the genome. Since enhancers are regulatory sequences, they have been typically searched in intergenic regions (**intergenic enhancers**), which represent approximately 98–99% of the genome [48]. However, studies have shown that there are also **intragenic enhancers**, which can be located in intronic (**intronic enhancers**) and exonic sequences (**exonic enhnacers or eExons**) and modulate gene expression, e.g. by acting as alternative promoters [49,50]. Furthermore, there are intragenic enhancers which are capable of recruiting TF that increase the recruitment of RNAPII and other GTF to the promoter of the gene itself [51] or to other genes [52]. On the other hand, other studies showed that the presence of enhancers in intragenic sequences can also attenuate the expression in the gene itself, possibly by an interference between the elongation of the gene transcript and the transcription of the enhancer [52]. Thus, the transcriptionally active intragenic enhancers could have disparate functions and improve transcription of one or more distal genes while limiting transcription of the gene itself (double function, as enhancer and silencer).

Enhancers by clustering: Enhancers can be found proximate and linked to the same regulatory process. For this reason, the idea was to cluster and merge these sequences to consider them as a single regulatory elements.

The term **distributed enhancers** has been coined to represent that multiple enhancers can regulate the same gene and to eliminate the controversy about the classification of enhancers as primary or secondary based on the distance to the target gene [11]. However, currently we know that multiple enhancers can act on the same gene and that one enhancer can act on more than one gene [9,53,54], so a deeper investigation of the regulatory profile may result in a generalization of this distributed or collaborative property that eliminates this concept as a subtype of enhancer. In general, it is common that when one enhancer explains a large part of an expression pattern is found, the search for more sequences is not carried out, so the number of genes with distributed enhancers could be high.

Superenhancers (SE) are the type of cluster of merged enhancers located proximally in the genome, typically within 12.5 kb, and have unusually high levels of master TF, RNAPII, cofactors like the mediator and integrator complex and other enhancer-associated features, such as H3K27ac and H3K4me1. They are also associated with high transcription values of eRNA and associated target genes, but also with a higher frequency of chromatin interactions with respect to individual enhancers [55–61]. In addition, the enhancers that compose the SE (**constituent enhancers**) are usually linked to the same regulatory process. In the identification of SE, first the enhancers located within a distance of 12.5 kb are joined and then, those sequences that exceed the inflection point in the signal level plot [55] are selected. After this process, enhancers not classified as SE are called **typical enhancers**.

SE are the most studied type of enhancer due to their relationship with important regulatory processes such as cell development and cell identity. This relevance derives from their ability to bind master transcription factors and their association with the expression of pluripotent genes and tissue-specific genes [55–57]. These characteristics and relevance also result in the association of these sequences with the development of diseases like cancer, and their consideration as therapeutic targets [62–64,7,65]. Other studies also suggest other aspects: individual enhancers may have a role equivalent to SE [66], a cooperative profile in SE sequences, not all elements have to act at the same time and under the same conditions, and a hierarchical structure within SE [66,13,67,42].

Some studies have tried to dissect this hierarchy in SE to determine which sequences are essential and which are more dispensable [61]. In turn, this organization has also included new terms or subtypes: **hierarchical and non-hierarchical SE.** Non-hierarchical SE have enhancers with similar contact frequencies and, therefore, they are more homogeneous. The hierarchical SE are more heterogeneous and have some enhancers with a higher frequency of interactions (**hub enhancers**) than the rest of the constituent enhancers (**non-hub enhancers**). Hub enhancers share similar histone marks with the non-hub enhancers, but have more CTCF and cohesin binding sites. Therefore, it has been suggested that hub enhancers act as organizational centers within the SE, that coordinate contacts with the rest of the non-hub enhancers and with other distal regulatory elements [61]. In addition, hub enhancers are more associated with SE function and disease-related variants, so their manipulation or deletion has demonstrated deep effects on gene activation and local chromatin state [61].

Enhancers by sequence length: This classification is based on the length of the sequence of the enhancer. 800 bp is the average length identified, **stretch enhancers** were defined as those longer than ⩾3 kb [68], otherwise they are **standard enhancers**. These sequences were also associated with genes with higher expression levels and with cell type-specific genes [69].

Enhancers by sequence overlap: This classification is based on the existence of overlap in the position and sequence of the enhancers. The classifications by sequence length and by clustering have identified SE and stretch enhancers. Around 85% of the SE overlap with stretch enhancers, SE being the set of stretch enhancers with higher activity and with higher enrichment values in analyzed tags. The overlapping sequences of SE and stretch enhancers were designated as **super-stretch enhancers** [68].

Other enhancers are overlapped with known promoters and exhibit enhancer activity because they can interact with other promoters [70,71] and have bidirectional transcription [72]. With the improvement in gene editing systems and high-throughput reporter assays, distal enhancer activity in promoter sequences was verified [73–77]. **Epromoters** is the term used for these sequences and some works consider that 2–3% of promoters are of this type [76].

We also find overlapping enhancers with Locus Control Regions (LCR), because they are structures composed of different regulatory modules that can include enhancers [78,79]. However, they have also mapped onto sequences annotated as silencers, which have the opposite biological function to enhancers. This is because regulatory sequences sometimes have a bifunctional function depending on cellular context [80,81]. However, a special nomenclature is not used for these sequences. We represent these cases in the Fig. 3 as **LCR enhancer and silencer/enhancer**.

Enhancers by chromatin profile: Different terminologies and classifications have been established according to the chromatin profile [26,82–86]. This classification is one of the most important because the activity of enhancers is not usually measured directly, but their activity is inferred from the chromatin profile of the sequences. As the activity of enhancers varies depending on the biological context, this classification is directly linked to the biological sample used.

Previously, the community thought that when a cell had completed its terminal differentiation, the regulatory repertoire was established and maintained by lineage-specific TF. Thus, internal or external stimuli could not change the regulatory pool, but acted within it through the cooperation with the master TF that were already bound to the sequences. However, cellular plasticity has shown that stimuli can create new functional properties through the activation of regulatory sequences that are not pre-established by the cell lineage, even to the point of defining a new cellular subtype [85]. In response to stimuli, some enhancers without histone masks characteristic and without bound TF (**inactive enhancers**) can recruit master TF that provide chromatin accessibility and allow the acquisition of a chromatin profile associated with enhancer activity, such as H3K4me1 and H3K27ac. These inactive enhancers without marks that can get activated after a stimulus are called **latent enhancers** [85]. After the loss of the stimulus, some chromatin marks may be lost, like acetylation and TF binding, as well as regulatory activity, but the sequences can retain marks like H3K4me1. Subsequently, when cells receive a stimulus again, the sequences can be re-stimulated with a faster and stronger response [85].

Sequences with H3K4me1 were initially considered as enhancers with little or no activity, but predisposed to acquire acetylation and transition to a more active state. These sequences were named as **primed/poised enhancers** [85,26]. However, studies have shown that the activity of sequences labeled with H3K4me1 is not lower because of lacking H3K27ac, but they could contribute to expression in a similar or even superior way to the enhancers labeled with acetylation and called as **active enhancers** [87]. On the other hand, it was also observed that these predisposed enhancers can also present marks associated with the silencing, like H3K27me3, H3K9me3 and PRC2 binding, but also P300 (associated with activity) [88,89,83,90,91]. These bivalent sequences have been found close to inactive genes that are important during development and can be activated during differentiation by deletion of H3K27me3 and gain of H3K27ac. In addition, chromosome conformation capture assays have shown that these sequences can be detected interacting physically with their target genes through the PRC2 complex even before activation [90,20,92,93].

The observed variability has led to different subdivisions of this set, but also to different nomenclatures due to the lack of consensus and controlled vocabulary. Some studies classify and name the enhancers without tags as inactive enhancers, the enhancers H3K4me1+ as **primed enhancers** or **intermediate enhancers**, the enhancers with marks like H3K27me3 and PRC2 a subset of this primed enhancers are called **poised enhancers** or **bivalent enhancers**, and the enhancers H3K4me1+ and H3K27ac+ as active enhancers [26,83,94,95]. Other classifications and nomenclatures are common in the chromatin state annotation of different studies and projects, like Roadmap epigenomics, which have generated datasets widely used in other research and public sources [96,82,86,97]. In these annotations we find categories such as:

- Strong and weak enhancers [82], similar to a classification reduced to active and primed enhancers.
- Genetic enhancers, enhancers and bivalent enhancers [96], equivalent to active, primed and poised enhancers.
- Enhancer, permissive regulatory region and bivalent enhancer [95], also equivalent to active, primed and poised enhancers.
- Active, poised, repressed and inactive enhancers, similar to active, primed, poised and inactive enhancers, respectively [98].

It is important to note that, although some labels are closer to a classification by activity, these labels correspond to a classification by chromatin profile, because they are assigned on the basis of chromatin properties. These properties are correlated with activity, but activity must be confirmed experimentally and it has also been shown that these tags are not strictly necessary to identify activity in enhancer sequences [35,99,87].

Enhancers by transcription: Transcription has been observed in some enhancers, mainly those considered to be active [29,53,100,101]. Therefore, enhancers can also be classified into **transcribed enhancers (T-Enh)** and **non-transcribed enhancers (NT-Enh)**. This transcription is initiated in the NFR of the enhancers and is mainly bidirectional, although eRNAs with structural heterogeneity and different combinations of properties have been observed: unidirectional, bidirectional, polyadenylated and non-polyadenyadenylated [27,29,102,103]. Within this heterogeneity, the most usual subdivision is between enhancers with unidirectional (**1D-Enh**) and bidirectional transcription (**2D-Enh**), although a recent single-cell CAGE sequencing suggests that the directionality of transcription could be more complex than unidirectional or bidirectional in absolute terms [29,102]. Different functions have been proposed for the transcripts [17,29,103–105], such as formation and stabilisation of the chromatin loop [106], liberation of the transcriptional pause [107] or increasing the occupancy of TF and coactivators in enhancer sequences [108]. Even eRNA function can be different depending on the transcribed strand [109].

2.1.2. Methodologies to generate evidence

The variety of enhancer features also results in a variety of different experimental approaches for their detection, which provide different levels of evidence. Comparative studies of enhancers obtained by different methodologies have shown that, currently, there is no preferential method for the detection of enhancers [36]. Each methodology provides a set of candidate sequences because they have genomic characteristics that make them potential enhancers [24]. Nevertheless, each of them only provide a partial view of the profile of enhancers due to technical limitations, and because most techniques identify sequences indirectly, through characteristics correlated with enhancer identity and activity. For this reason, the inclusion of supporting methodologies is essential in an enhancer representation model, because they provide the level of evidence to candidate sequences that need to be verified. Fig. 4 shows the two groups of evidence generation methodologies identified in the literature according to the identification strategy used, namely, based on chromatin characteristics and based on reporter assays. Next, we describe both types.

Evidence based on chromatin characteristics: The objective is to identify sequences according to chromatin properties correlated to enhancers. The development of high-throughput genomic methods has been fundamental to capture these properties, while bioinformatics tools have allowed us to analyze, to search patterns in the data and to generate models to classify sequences according to chromatin properties [110–112,86,96]. Therefore, these properties are the evidence supporting the enhancers and we distinguish different approaches or levels of evidence which we cover below.

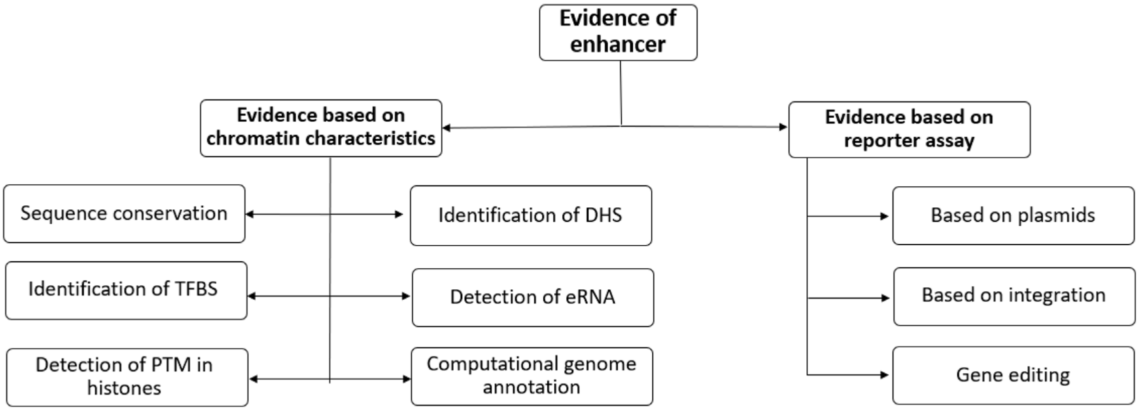

Fig. 4. Enhancers can be identified through different methodologies, which follow a certain strategy or approach that provides the level of evidence for the sequence. We can distinguish two main types of evidence. Based on chromatin features: they appeal to sequence features, thus correlated properties that are not a direct measure of enhancer activity. Reporter-based: these measure enhancer activity directly, but the interpretation of the results can be complex.

- **Detection of sequence conservation** involves finding conserved sequences between species and over time. It was one of the first approaches used for the identification of enhancers and followed in the early repositories. The detection of conserved sequences has been successful in the discovery of enhancers involved in biological processes of high importance in most organisms, like sequences active during early development, where enhancers have represented almost 50% of the highly conserved sequences analyzed [113]. However, this approach presents problems in the detection of sequences with species specificity, where evolutionary conservation is lower [3]. In addition, studies show that enhancers have different levels of evolutionary conservation when they are obtained by different identification methods. All of them show a higher overlap with conserved elements than randomly, but the enhancers detected by eRNA transcription were the most conserved, while the enhancers obtained by PTM marks were the less conserved [36].
- The **identification of TFBS** is an approach based on the fact that the activity of many enhancers depend on their ability to bind TF and coactivators that, by different possible mechanisms, increasing the transcription of target genes [17,114]. For this purpose, ChIP-seq has been used to capture and sequence those DNA fragments bound to a specific TF, p300 and the mediator complex being the most commonly used [115,58]. However, this method may have low specificity because, in addition to non-specific binding, can capture those sequences that have complementary sequence and chromatin accessibility. This is also an antibody-dependent method with its associated problems [23]. Therefore, it is not a method typically used isolated for the identification of enhancers, rather it is used to generate chromatin profiles that integrate different chromatin properties in order to classify and predict sequences.
- The **identification of accessible chromatin (DHS)** is another approach used in the identification of enhancers. If many enhancers base their activity on their ability to bind TF and on the mechanisms of action of the eRNAs [17], these sequences must be accessible, at least during their period of activity. This approach has low specificity, because multiple non-enhancer sequences share this property [116,117]. Therefore, as for TFBS, the identification of DHS is generally used for chromatin profiling.
- The **detection of PTM in histones**, such as methylations and acetylations, has been widely used both as individual approach and for building chromatin profiles [26]. For this, chromatin immunoprecipitation followed by microarrays or sequencing is the technique most popular [118]. However, there is not strictly necessary and reliable methylation or acetylation mark to identify unambiguously enhancer sequences [35,99,87]. In the beginning, H3K4me3 was associated with promoter sequences and H3K4me1 with enhancer sequences, while dimethylations were observed in both types of sequences without a clear distinction [119]. H3K4me1 and H3K4me3 are not mutually exclusive marks in a genomic region, so the H3K4me1/H3K4me3 signal ratio has also been used [120,121]. However, this criterion was affected by the detection of enhancers with high H3K4me3 values [120,122] as well as enhancers without H3K4me1 [123–125]. On the other hand, H3K4me1 is not a mark capable of discerning enhancer activity. Acetylation analysis associated the H3K27ac mark with the activity of the enhancers, so this mark has been widely used in the identification of active enhancers [88,89]. However, other studies have also shown that H3K27ac is also not a strictly necessary mark for enhancer activity [126,87]. For these reasons, histone PTM are also used with DHS and TF binding data for the development of computational models that annotate chromatin following chromatin profiles.
- **Detection of eRNA**. Since transcription has been correlated with sequence activity, eRNA detection has been used to identify active enhancers [53,127–129]. The methods employed for the sequencing of these RNAs are varied. There are techniques that allow the detection of RNA already produced, either the full length of the sequence (e.g., flcDNA-seq) or the first nucleotides (e.g., CAGE and TSS-seq). Other techniques use the transcription rate (e.g., GRO-seq, PRO-seq or Start-seq). These nascent RNA sequencing techniques allow to measure transcript levels, so they have the advantage of quantifying RNA sequences which are not very stable and are rapidly degraded. This is the case of eRNAs and PROMPT sequences of promoters [99].
- **Computational genome annotation**. The development of algorithms able to work efficiently with large volumes of data has also made it possible to work with multiple experimental evidence rather than individual chromatin properties. These models or algorithms have presented the problem of enhancer identification from a computational point of view and their goal is to determine if a sequence can function as an enhancer or not according to a set of multiple types of data that provide a description of the sequence [130,131]. Therefore, the first step in these algorithms is the integration of different types of data that provide information about the sequences. Subsequently, these data are preprocessed (normalisation and scaling) to generate a feature vector that serves as input for an enhancer iden-

tification and analysis system, which is responsible to annotate the DNA regions based on these feature vectors. The computational models used have been developed following different computational strategies and we find supervised and unsupervised methods. Some tools developed include clustering algorithms, like K-means or bi-clustering; others use regression models, like least absolute shrinkage and selection operator (LASSO); probabilistic graphical models (PGMs), like Dynamic Bayesian Networks (DBNs) and Hidden Markov Models (HMM); or classification systems like artificial neural networks (ANNs), support vector machines (SVMs), random forests (RFs) and decision trees (DTs) [130].

Evidence based on reporter essays: The objective is to identify whether a sequence can increase the expression of a reporter under the control of a minimal promoter. With the development of high-throughput methods like MPRA and STARR-seq, this approach can now be used for massive sequence identification [132–134,24,135]. According to the reporter method used, the enhancer can drive the expression of: a given sequence, such as a barcode used as a reference; its own expression, through eRNA measurement; or the expression of an alternative reporter gene, like a fluorescent reporter. The vector used also varies the methodology.

- **Assays based on plasmids**. This is a simple approach with higher throughput, but is unable to replicate the complexity of gene regulation in chromosomes. Examples are episomal reporter assays, STARR-seq and MPRA.
- **Assays based on integration**. The integration into the genome is a complicated process and can be done randomly or in a guided manner. If random integration is chosen, the genomic context can be lost, while in guided integration the context can be maintained, but at the cost of low efficiency due to the limitation of the number of sequences that we can analyse in parallel. On the other hand, *in vivo* systems are more reliable than *in vitro*, although we should not ignore the technical limitations and problems that may arise from the genomic context, cell and organism specificity when we use model organisms [132,133].
- **Gene editing** with CRISPR-Cas9 technology and the use of guide RNAs (sgRNA) have also facilitated the endogenous manipulation of enhancers [136,66] and can help to screen sequences even at the TFBS level. The main advantage of this method is the possibility to work *in vivo* at a high scale, in a targeted manner and to maintain the local chromatin context by using targeted sgRNA libraries [137,138]. It also allows other molecules to be incorporated into the Cas9 nuclease, so many technical variants have been developed to activate and silence sequences [139,74,140]. However, the main problem is to be able to evaluate the causality and impact of the alterations of the enhancers on expression, as well as the scaling of the technique.

2.1.3. Enhancer-promoter interactions

The public sources often include information that enriches the knowledge about enhancers. Within this information, enhancer-promoter relationships are one of the most important because they inform about the potential regulatory role of the enhancer, either by regulating the transcription of protein-coding sequences or non-coding sequences, such as lncRNA or miRNA [141–144]. The distance between enhancers and promoters that interact can be very large, with a distance of one megabase or more, but they are usually closer [10,71]. The majority of enhancer-promoter interactions (EPI) are less than 200 kb and the most numerous are usually around 20–50 kb [71,20]. Chromatin conformation capture methods (3C and derived methodologies) and high-throughput microscopy techniques, such as fluorescence in situ hybridization (FISH) experiments, are experimental tools used to study the three-dimensional structure of chromatin and to determine the contacts between sequences [145,31,146,147,32–34]. Computational methodologies are also used to predict EPI [141,145].

Therefore, similar to enhancer identification, the strategies used in the different sources for EPI identification are varied (Fig. 5) and, despite advances in new technologies, comparison of different methods against a curated reference set has found that they still need to be improved [37]. The association methods can be classified in two groups: unsupervised machine learning and supervised machine learning [145].

Unsupervised methods: In unsupervised models, machine learning is based on a model that is adjusted according to the observations, so there is no *a priori* knowledge. Within this type we highlight the methods based on distance, correlations and scores.

- **Distance-based methods** link enhancers to target according to one or more distance functions. Linkage to the closest TSS has been the most widely used strategy and estimates have determined that approximately 40% of EPIs are established with the gene linked to this gene [53]. The association with the closest gene could happen for different reasons, such as a low specificity of the enhancer for its target sequence [20]. Other widely used distance-based strategies include: overlapping genes with the enhancer, proximal genes or genes within a distance window, flanking genes or closest genes on both sides of the sequence. Distance-based models can be useful for generating an initial list of possible target genes, but it is a method that can be inefficient because it does not consider other aspects like interactions at large distances, cooperatively between sequences or the specificity [13,45,9,148].
- **Correlation-based methods** evaluate the correlation of properties between pairs of sequences, such as correlation between eRNA transcription and gene transcription [53], or correlation between eQTL values and active chromatin marks [82]. A significant advantage of these methods is that they can identify multiple targets for one enhancer and measure quantitatively the power of the association [141]. In contrast, a major problem is the need for a large number of samples with sufficient quality for comparisons, because correlation methods assume that enhancer activity changes between conditions and between cells [141]. They are also very sensitive to outliers and can therefore generate a large number of false positive predictions, but some algorithms have been developed to deal with this outlier problem [149].
- **Score-based methods** integrate data of different types and each feature is associated with a quantitative value that is used to generate a total quantitative score to establish an association ranking for enhancer-gene interactions [141]. These methods have also been defined as Decomposition-based methods in other sources [145]. An advantage of these methods is that all possible interactions between sequences can be quantified, you can adjust the level of significance and allow different priorities for each of the features. However, the need to adjust weight values to the features is also one of the main problems, because this can be arbitrary.

Supervised methods: In supervised learning models, we start with a set of data that are true EPI, i.e. true positives and negatives, and we use them as a training set to find patterns and create models that can be a classifier or a regression model.

- **Classifiers** use the patterns found in genomic features to create a model that generates labels that are then applied to new datasets. As with correlation and scoring methods, the number and type of features used for the predictor can be highly variable,

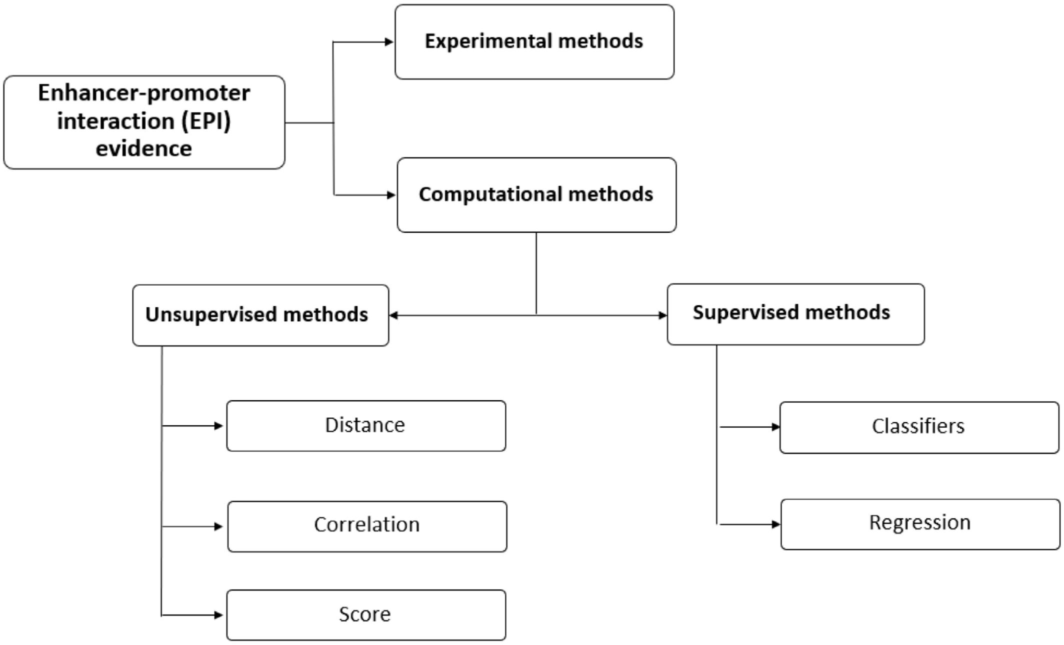

Fig. 5. Similar to the identification of enhancers, the determination of EPI can follow different strategies, which provide the level of evidence for the regulatory relationship. Two main groups are also distinguished. Experimental methods determine the relationship directly. Computational methods make predictions and can follow two main approaches. Supervised methods generate a model from a training set, while unsupervised methods lack this set.

and the potential of these models will depend on the data used for training and the quality of the variables [145,141]. The main advantage of these systems is that they can find hidden patterns in the data that are difficult to see when we have a large amount of data, while the main problem is that they are very dependent on the dataset used for training.

- **Regression-based methods** differ from classifiers primarily in their ability to quantify potential. These methods systematically evaluate the quantitative contributions of enhancers that can regulate a gene within a genomic window by exploiting a large number of genomic features that are assigned contribution weights. These methods work with the logic that multiple enhancers can regulate one gene. Therefore, a combinatorial approach is applied for sequence pairing [141]. This has the advantage of the cooperative effect of the enhancers under consideration. However, they have the disadvantage typical of supervised machine learning methods in terms of training data. Besides, it also requires the definition of a genomic window, which is set by a distance criterion, and a maximum number of enhancers to be considered around a TSS.

2.1.4. Other annotations of interest

In the annotation of enhancers, besides the sequence coordinates, we have emphasized the importance of the different types, which in the case of a chromatin profile classification depends on the biological sample. That is because the activity of the enhancer sequences can be cell-specific, but also stimulus-dependent [13,9,3,150]. Therefore, the chromatin profile of the sequences also varies between biological samples and, consequently, the type. All cells in an organism have the same genetic information and therefore all cells have the same enhancers. Thus, simply identifying enhancers in the genome tells us nothing about possible enhancer activity and cell-to-cell variability. In this way, the annotation of the type of enhancer and the biological source is essential for a model of enhancer representation that aims to study gene regulation.

We have also pointed out the importance of the methods which support the sequences as well as possible target genes, because they are the target of regulation, and their corresponding EPI predictor methods, because they also support the prediction. However, the representation model can be enriched with other annotations of interest that increase its value. The activity of enhancers derives mainly from their ability to bind TFs and to generate functional eRNAs [17]. Alteration of enhancer activity can also contribute to the disruption of regulatory networks and the development of diseases [4]. Therefore, including information about TFBS, eRNA transcription, mutations and linking enhancers to biological networks and diseases enriches the value of the model, also the annotation of chromatin profiles, mechanisms of action and other useful information.

2.2. Databases

Databases store the information generated by the scientific community about enhancers. The volume of sequences obtained by the different research efforts varies according to the identification method used, because most methods use different correlated features, or combinations of correlated features, to identify enhancer sequences in the genome. The consequence is that the results can differ in several orders of magnitude [36] and this variability is also reflected in the repositories, because each source identifies and stores sequences according to different criteria and data inputs. Thus, while the RefSeq reference genome GRCh38.p13 (release 109.20211119) contains around 5,000 enhancers, FANTOM5 contains around 50,000 sequences [53] and SCREEN ENCODE around 1 million [151].

For our study, we have selected 25 publicly available and accessible repositories specialized in identifying and annotating human enhancer sequences and which annotate, at least, the coordinates of the sequences (see Table 1). The data was collected in February 2022.

Table 1
Brief description of the databases included in this study.

Repository	Focus	Short Description
CancerEnD	Diseases	Set of enhancers for TCGA cancer types
dbInDel	Mutations	Enhancer-associated insertion and deletion variants
dbSUPER	SE general annotation	Super-enhancers archive
ENdb	Diseases	A manually curated database of experimentally supported enhancers for human and mouse
EnDisease 2.0	Diseases	A manually curated database for enhancer-disease associations
EnhancerAtlas 2.0	General annotation	General annotation of enhancers in different human biosamples and other species
EnhancerDB	General annotation	General annotation of enhancers in different human biosamples
EnhFFL	Feed-forward loops (FFL) with enhancers	A database of enhancer mediated feed-forward loops for human and mouse
Ensembl Regulatory Build v105	General annotation	Set of regions of the genome that probably are involved in gene regulation
ETph	Pig-human homology	General enhancers and their targets in pig and human
FANTOM5	Transcribed enhancers	Transcription-capable enhancers
FOCS	EPI	Method for inferring an extended enhancer-promoter and predicted set
GeneHancer 4.8 (UCSC)	General annotation	Integration of enhancer sequences to generate a consensus set
HACER	Transcribed enhancers	Transcription-capable enhancers
HEDD	Diseases	Human enhancers with a focus on their links to diseases
HeRA	Transcribed enhancers	Transcription-capable enhancers
RAEdb	Enhancers identified by reporter assays	Enhancers identified by high-throughput reporter assays
SCREEN V3	General annotation	Set of regions of the genome that probably are involved in gene regulation
Roadmap epigenomics	General annotation	Genome annotation in states
SEA 3.0	SE general annotation	Super-enhancers archive
SEanalysis	Biological networks with SE	Super-enhancers associated with regulatory networks
SEdb 1.03	SE general annotation	Super-enhancers archive
RefSeq GRCh38.p13	General annotation	Annotation of functional elements in the reference genome
TiED	General annotation	Identification and annotation of active and transcribed enhancers in 10 tissues
VISTA Enhancer	General annotation	Validated enhancers with transgenic mice

3. Results

In this section we describe how the selected databases cover the information about enhancers included in our model.

3.1. Types of enhancers

Table 2 shows that the majority of the databases do not cover the types of enhancers, but annotate the sequences as general enhancers (see bar plot in Fig. 6A). Therefore, both the type of enhancer and its possible activity profile have to be inferred mainly from the methodology used for sequence identification, through an analysis or by other means. Due to their relevance, the most covered enhancer type in the repositories are SE and transcribed enhancers, although the constituent enhancers are not always included. dbSUPER [152], ENdb [153], SEA [154], SEdb [155], SEanalysis [156] and EnhFFL [157] are repositories which contain SE. RAEdb [158] is the only source that covers epromoters, while SCREEN [151] distinguishes between proximal and distal enhancers according to their distance to the nearest TSS (2 kb limit). On the other hand, according to chromatin profile we find mainly two sources: Ensembl [98] and Roadmap [96]. The first distinguishes between Active, Poised, Repressed, Inactive and NA. The second does it between Genetic enhancers, Enhancers and Bivalent enhancers.

Table 2
Type of enhancers hosted by each database. The types of enhancers not included in this table are not covered by any database included in this study.

Enhancer types according to model	Repositories
Enhancers (without classification)	CancerEnD, dbInDel, ENdb, EnDisease 2.0, EnhancerAtlas 2.0, EnhancerDB, Etph, FOCS, GeneHancer 4.8, HEDD, RAEdb, RefSeq GRCh38.p13, VISTA Enhancer
Super-enhancers	dbSUPER, ENdb, EnhFFL, SEA 3.0, SEanalysis, SEdb
Typical enhancers	EnhFFL, SEA 3.0, SEanalysis, SEdb
Constituent enhancers	dbSUPER, SEanalysis, SEdb
Epromoters	RAEdb
Proximal enhancers	SCREEN V3
Distal enhancers	SCREEN V3
Active enhancers	Ensembl Regulatory Build v105, Roadmap, TiED
Primed enhancers	Ensembl Regulatory Build v105, Roadmap
Poised enhancers	Ensembl Regulatory Build v105, Roadmap
Inactive enhancers	Ensembl Regulatory Build v105
Transcribed enhancers	FANTOM5, HACER, HeRA, TiED

Furthermore, while Ensembl first annotates the consensus enhancer sequence in the genome and then profiles the type of enhancer according to the biological sample, the other repositories usually annotate the sequences by biological sample, without finding a reference sequence. Therefore, the amount of enhancers in databases is usually very high, because each biological sample annotates enhancers that may coincide with those of another biological sample or overlap and differ in sequence boundaries. Therefore, this amount is reduced when we obtain sequences with unique coordinates, and could be further reduced if we search for consensus sequences from overlapping sequences that differ at the boundaries.

3.2. Methodologies to generate evidence

The first aspect studied was the origin of the data in the resources (see supplementary material) and whether the resources that included data from different databases perform an integration of the data or preserve original sequences. First, we found repositories that store data from their own study, such as FANTOM5, and repositories that integrate data from different sources, such as ENdb. In turn, these integrative repositories can enrich the information with their own contributions or include new sequences. Regarding the sequences, we can find examples of both situations. On the one hand, we find repositories that compile enhancers from different sources to generate a new set, such as Genehancer [159], EnhancerAtlas [160] or HEDD [161]. On the other hand, there are repositories that preserve the original sequences, such as dbSUPER

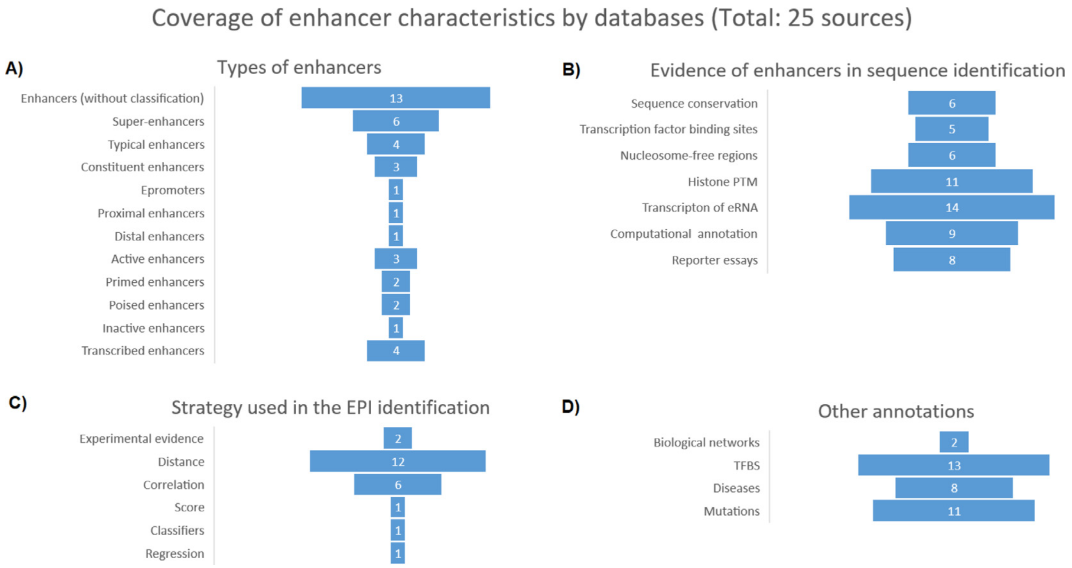

Fig. 6. Coverage of the different items in the 25 biological databases analyzed with information about human enhancers.

[152], EnDisease [162] and ENdb [153]. In addition, in both cases, the sequences do not usually provide information or cross-references to the original source records, which makes it difficult to contrast and integrate the information and to follow a historical record of the evolution of the data. Table 3 summarizes the results obtained for each identification method and that are described next. In quantitative terms, repositories that integrate data and generate new sequences are more frequent, while the most used evidence use strategies based on chromatin properties. Fig. 6B shows that 14 of the 25 databases analyzed have enhancers with eRNA transcription evidence. This is because FANTOM5 is a repository widely used by other repositories as a source (13 repositories use the FANTOM5 sequences, see supplementary). On the other hand, by volume of data, strategies based on PTM and computational annotation provide the highest number of sequences (see supplementary material).

3.2.1. Evidence based on chromatin characteristics

Detection of sequence conservation: This is the case of VISTA Enhancer, which selects candidate sequences by sequence conservation and subsequently validates them by reporter gene assays in mouse embryos [163]. It is also used by GeneHancer for enhancer confidence scoring [159].

Identification of sequences that bind TF: This is the case of the methodology used by sources like dbSUPER [152], which contains SE identified by Med1 and BRD4, and sequences included in EnhancerAtlas [160]. It is also a methodology used to generate genome annotations in repositories such as SCREEN ENCODE [151], Ensembl [98] or RoadMap [96] and, therefore, by the sources that use these sequences (see supplementary material). In addition, many repositories enrich their enhancers with information about TFBS obtained from ChIP-seq experiments or by computational prediction, because this is relevant information in the study of

Table 3
Databases classified by the experimental evidence supporting the sequences that they contain.

Repository	Seq conservation	TFBS	NFR/DHS	PTM	eRNA	Computational annot.	Reporter essays
CancerEnD					X		
dbInDel				X			
dbSUPER		X		X			
ENdb	X	X	X	X	X		X
EnDisease 2.0							
EnhancerAtlas 2.0		X	X	X	X		X
EnhancerDB	X		X	X	X		X
EnhFFL		X	X	X			
Ensembl Regulatory Build v105	X				X	X	X
ETph					X	X	
FANTOM5					X		
FOCS			X	X			
GeneHancer 4.8 (UCSC)	X				X	X	X
HACER					X		
HEDD					X	X	
HeRA					X	X	
RAEdb							X
Roadmap epigenomics						X	
SCREEN V3						X	
SEA 3.0		X		X			
SEanalysis				X	X	X	
SEdb 1.05				X	X	X	
RefSeq GRCh38.p13	X						X
TiED			X	X	X		
VISTA Enhancer	X						X

enhancers and regulatory networks. However, this enrichment information is not always available for download (see supplementary material).

Identification of DHS: The identification of DHS is used more for chromatin profiling to identify enhancers. It is also a useful type of information to enrich sequence information, like is done in EnDisease source [162].

Detection of PTM in histones: The dbSUPER is an example of source that used PTM for the identification of SE, specifically the H3K27ac signal [152]. This mark is also used in other sources such as SEdb [155], SEanalysis [156], SEA [154], dbInDel [164] and EnhFFL [157]. On the other hand, EnhancerDB identified enhancers using high levels of H3K27ac and H3K4me1 and low levels of H3K4me3 [165]. Furthermore, histone PTM are also used with DHS and TF binding data for the development of computational models that annotate chromatin following chromatin profiles.

Detection of eRNA: The CAGE technique was the methodology used, for example, by the FANTOM5 consortium [53], so it is the technique that provides the level of evidence for these sequences. In addition, the dataset obtained by FANTOM5 has been widely used by other sources, both for sequence integration purposes and to add transcript enrichment. This is the case of repositories like CancerEnD [166], HeRA [167], FOCS [168], Ensembl Regulatory Build [98], EnhancerDB [165], HACER [169], SEdb [155], SEanalysis [156], EnhancerAtlas [160], GeneHancer [159] and TiED [101]. On the other hand, GRO-seq and PRO-seq technologies were used for the identification of enhancers in the HACER source [169].

Computational genome annotation: ENCODE, Roadmap and Ensembl are examples of repositories that follow this approach of computational genome annotation through integration of different experimental evidence, mainly PTM, DHS and TF binding [151,98,96].

3.2.2. Reporter-based methods

Reporter-based assays are not widely represented in the enhancer databases. We summarize next the results for this type.

Assays based on plasmids: ENdb source collects enhancers from the literature, some of which have reporter gene assays of this type as evidence [153].

Assays based on integration: After the identification of candidate enhancers by evolutionary conservation, the VISTA Enhancer repository validated the sequences using transgenic mice by reporter gene assays.

STARR-seq and MPRA high-throughput methodologies: These were used in the identification of RAEdb sequences [158], so EnhancerAtlas also contains information of this type because it integrates enhancers from this database [160]. The ENdb source collects enhancers from the literature, some of which have reporter gene assays as evidence [153].

3.3. Enhancer-promoter interactions

We have distinguished between experimental methods and computational predictions. Similar to enhancer identification, the strategies used in the different databases for EPI identification are varied (Table 4). It is important to note that repositories usually include both coding and non-coding sequences such as miRNA or lncRNA as target genes. HACER [169], GeneHancer [159] and ENdb [153] are sources of enhancers that incorporate data from 3C experiments and derivatives to annotate and predict potential target genes. Only ENdb contains EPI with evidences based on reporter assays and gene editing, which are the most commonly used experimental techniques to verify enhancer-gene regulation [24,141,145].

Regarding computational methods, distance-based methods are still the most used by biological databases (12/25), followed by correlation-based methods (6/25) (see Fig. 6C and Table 4). We can also note that not all databases include EPI (see supplementary material).

3.3.1. Unsupervised methods

Distance-based methods: Some sources annotate the closest genes, like SEA [170], VISTA Enhancer [163] or dbInDel [164]. Others use a window that varies in size according to the source. In EnhancerDB this window is ±100 kb [165], while in dbSUPER it is ±50 kb [152]. There are also sources that use a combination of distances. SEdb [155] and SEanalysis [156] include the strategies of nearest active gene, genes overlapping with the enhancer, proximal genes and results obtained by the Lasso [171] and PreSTIGE [172] algorithms. In the case of HACER [169], the closest gene and genes within a distance of 50 kb are included.

Correlation-based methods: FANTOM5 used correlation between eRNA and gene transcription to link enhancer and genes within a 500 kb window [53]. This set of associations established by FANTOM5 has been used to enrich other repositories, such as HEDD [161] and HACER [169], but also as part of other models. This is the case of the scoring system used by GeneHancer [159]. On the other hand, Roadmap [96] used a method based on correlation between eQTL values and active chromatin marks.

Table 4
Experimental approach used in the identification of EPI, which constitute the evidence of the regulatory relationship between sequences.

Repository	Experimental evidence	Distance	Correlation	Score	Supervised method
CancerEnD		X			
dbInDel		X			
dbSUPER		X			
ENdb	X				
EnhancerAtlas 2.0					X
EnhancerDB		X			
EnhFFL		X			
ETph		X			
FANTOM 5			X		
FOCS					X
GeneHancer 4.8 (UCSC)				X	
HACER	X	X	X		
HEDD			X		
HeRA			X		
SEA 3.0		X			
SEanalysis		X	X		
SEdb 1.03		X	X		
TiED		X			
VistaEnhancer		X			

Score-based methods: The GeneHancer repository is one example of a source that uses this system (based on eQTLs, CHi-C, eRNA co-expression, TF co-expression and distance) [159].

3.3.2. Supervised methods

Not many databases register EPI detected by applying supervised methods. The EAGLE algorithm, a classifier, was used by the EnhancerAtlas database [160], while FOCS is an example of a regression-based method that uses ordinary least squares to predict promoter activity as a function of k nearby enhancers within a window of ±500 kb [168].

3.4. Other annotations of interest

Our supplementary data file includes the main annotations included in each of the public sources analyzed with information about human enhancer sequences. However, most of them are available in the web version of the databases, but not for downloading. Fig. 6D shows that, statistically, the different annotations are covered by less than half of the databases analyzed.

4. Discussion

Enhancer are distal regulatory sequences that have been shown to be able to modulate gene expression, even over large distances, and to be fundamental in important regulatory processes such as development, cell identity, but also in pathologies that have been termed enhancerophaties [4]. Enhancers do not have a homogeneous profile, but there is a great diversity even between different tissues due to cell specificity [13]. For this reason, there are also different methodologies for the identification of enhancers that provide a partial view of the regulatory landscape. Identifying the relationship between genes and enhancers is not a simple task and different approaches have been used. All this variability of information has been transferred to the different databases.

This study has included 25 publicly available databases. There are more resources about enhancers in the literature, but they were excluded because of unavailability (e.g., DENdb, DiseaseEnhancer and SELER), not containing human data (e.g. AnimaleRNAdb and Zenbase) or containing non-specific sequences (e.g., PReMod and UCNE). The current situation is that there is no central repository, that each database has a different model and there is no cross-referencing between these databases. This makes the collection of information about enhancer sequences difficult and justifies the need for the analysis carried out in this work, which has been done based on a model for enhancer sequences extracted from literature. Next, we describe the major findings, gaps and challenges that can be drawn from our work.

4.1. Findings

These are the major findings drawn from our research. None of the existing databases can be considered the main entry point when searching for information about enhancers, since no database includes every type of information. The choice of database (s) will depend on the requirements and goals of our study. Given that the databases do not share a unified model, they are not interoperable, which makes it difficult to combine the information from the different resources.

The classification of enhancers is poorly covered in the databases. The classification into SE and typical enhancers is the most popular in databases, but the majority of repositories annotate the sequences in a general way by biosamples. SCREEN, SEdb, HACER and EnhancerAtlas exhibit the largest diversity in biological sources. We highlight the Ensembl annotation, because annotates the enhancers in the genome and subsequently classifies them according to their chromatin profile, a classification that allows us to estimate the activity of the enhancers in each biosample. However, this type of annotation could be expanded to cover the different types of enhancers.

Regarding the identification of enhancers, each database includes enhancers identified by different methodologies, providing a partial view of the regulatory landscape. For this, the strategy followed by EnhancerAtlas is of great interest, because it integrates enhancers obtained by different approaches, which can provide a broader view of the current knowledge. However, in the generation of the consensus set, the database does not provide the original sequences that produce the new enhancers and their methodologies, so we cannot keep track of the historical record and support for the sequence prediction.

The majority of databases use basic distance strategies for enhancer-promoter interactions. Since there is no preferred prediction method according to the literature, it is positive to have predictions developed by different strategies. Therefore, in a similar way to enhancer identification, we should highlight the annotation of repositories such as HACER, because includes EPI identified by different strategies.

With respect to the other annotations of interest, only the SEanalysis and EnhFFL databases includes biological networks. The number of repositories about enhancers-diseases relationship is also small, as well as the volume of information they contain (see Fig. 6D and the supplementary material). Therefore, the study of the influence of enhancers on diseases is limited with these specialized databases. However, other databases without a focus on diseases contain biological biosamples associated with pathologies. Thus, against this situation, the comparative study of the enhancer profile between pathological and healthy biosamples may be an alternative.

Enrichments related to other data of interest such as TFBS or mutations also vary between repositories. Moreover, many of these enrichments are only available in the web version of the databases and are not available for download. This complicates the use of this information, because the repositories also do not usually have APIs to program queries.

4.2. Challenges and future directions

Next, we describe the main challenges in the field that we have identified due to our study. We also propose research directions of interest in this area.

4.2.1. Identification of enhancers

The study of enhancers also confronts other challenges associated with the identification of sequences, their target genes and the validation of candidates. In addition to the limitations associated with the experimental and computational techniques used, these challenges derive fundamentally from the identification and association of genes by indirect methods, because most of the methods use correlated properties that are not a direct measure and that only offer a partial view of the enhancer profile, as well as false predictions. Therefore, the validation of results also becomes a fundamental pillar that is also limited by the interpretation of the results, because the specificity of the regulatory sequences and their genomic context-dependent activity make this task difficult.

In this context, progress in high-throughput reporter assays such as MPRA, STARR-seq and gene editing with CRISPR-Cas9 would be a potential tool for a massive screening of candidate enhancers to validate their role as regulatory elements. In the meantime, the enhancer sequences hosted in the different resources should be considered candidate sequences, whose experimental validation is pending. The contribution of the scientific community is essential to submit scientific results to the databases to update existing knowledge, as well as keeping the databases updated and working, avoiding the obsolescence of the content and/or the shutdown of the databases.

On the other hand, novel software to identify enhancer sequences is being developed [173,174]. Comparative studies of algorithms and revisions about these tools have been previously elaborated in other works [130,175,176], although a more recent in-depth review regarding this issue would be of interest. In the supplementary material we have included the main algorithms that have been used to identify the enhancers provided in each repository. It is remarkable that the majority of these software are not specific tools for enhancer detection, but are common tools for peak identification, alignment and sequence processing due to the approach/strategy used for the detection of these sequences (Fig. 4 and Table 3). Moreover, the databases do not annotate this information, but users must check the original papers for more information on this issue. In addition, some papers report the identification process in the methodology but do not go into the software used, or use their own code, so the inclusion of this data is difficult and may not be complete. Therefore, the annotation of these tools that provide evidence is also an aspect that databases should improve and include in future repositories.

4.2.2. Underrepresented concepts in biological databases

With the exception of SE, the types of enhancers are underrepresented. In this case, the annotation of SE also needs to be improved, because the constitutive enhancers that compose the sequence are often not included. The classification by chromatin profile is particularly interesting, because chromatin marks are correlated with enhancer activity. Since the majority of repositories do not label the type of enhancer, the type has to be inferred based on the methodology used in the identification of the sequences and, therefore, the activity of the enhancers has also to be inferred. However, this annotation is not usually included in the databases either, but must be obtained from reading the corresponding article, a situation which becomes more complicated when the repository uses different experimental approaches, because the experimental evidence of the sequences may be lost. Therefore, an interesting area of further research is to explore the diversity of those sequences and their different profiles, which would increase the knowledge about the different typologies of enhancers. Also, the annotation of the type of evidence that supports the sequences is usually missing in the databases. That type of evidence is needed to properly report the validity of the data. The use of resources such as the Evidence Ontology should be considered [177].

A similar situation is found in the integration of sequences. Many repositories integrate information from different sources (see supplementary material), either to generate a new dataset, to increase the volume of data in the repository, or to use these sequences to add new useful information. However, databases lack cross-references between sources and do not keep the identifiers used in the reference sources. This representation complicates the identification of sequences and the monitoring of the evolution of information. This is an important aspect because each database has a different approach and does not provide all the annotations that may be of interest. It is therefore necessary to consult sequences in different repositories and the lack of linking between sources and the use of common identifiers makes this task difficult. Therefore, future work on the definition of community standards identifying, for instance, the minimal amount of information [178] that should be reported in the databases and how to represent that minimal information and those cross-references would help to have more homogeneous datasets and to facilitate link discovery. In this context, our model is offered as a tool for the representation and structuring of knowledge, as well as the use of identifiers instead of variable string variables between sources.

Biological networks are the least covered aspect of enhancer annotations. The association between enhancers and diseases is also an under-covered aspect. However, repositories often contain both healthy and pathological biosamples, so the information present can be explored for comparative studies. It is important to note that, although the annotation of biological samples is often carried out, the system of representation is suboptimal, strings are annotated instead of instances corresponding to a knowledge model. Therefore, choosing biosamples of interest between hundreds of possibilities in the repositories can be a complex task, because it is an unstructured annotation. In this task, the annotation of ontology instances can help, as it would allow to obtain the samples belonging to a level of granularity level of interest to the user.

4.2.3. Formal knowledge model for enhancers

The variety of characteristics detected in enhancers has led to a lack of consensus on the definition of these sequences [12] and the proliferation of different subtypes of enhancers described in the literature [26], some of which overlap between them and make the understanding of the regulatory landscape complex. The representation of enhancers by biological sample in the repositories also contributes to this problem, because millions of overlapping sequences have been generated that vary in their boundaries.

In this article, we have provided a model that captures relevant information about enhancer sequences. However, that kind of model should evolve towards a knowledge model. Formalizing enhancer related knowledge in the form of an ontology would contribute to eliminate controversy, duplicity and to have a consensus. Ontologies are a useful tool both for structuring information and for its representation and are used in Life Sciences, the Gene Ontology being the most successful example of biological ontology [179]. That enhancer-related ontology would be the knowledge reference that would facilitate the comprehension and appropriate transmission of scientific knowledge. Currently, the Sequence Ontology (SO) [180] is the most relevant ontology about features and attributes of biological sequences. SO includes the enhancer class (SO:0000165), but does not contain the most common subtypes of the literature. The SO have recently been extended with new terms related to gene regulation as part of the collaborative research carried out in the GREEKC consortium [181]. More concretely, the terminology related gene expression has been updated in the cis-regulatory module (CRM) [182]. A similar effort should be pursued in order to incorporate in the Sequence Ontology the terminology related to enhancers included in our model and extracted from the literature.

4.2.4. Integrated data exploitation

In a search for information about enhancer sequences located in a region or that control the regulation of a certain gene, the current database landscape requires the user to query a wide variety of biological databases that are not interoperable with each other, which means that they cannot easily exchange information and that their information cannot be easily combined. The search tools provided by web portals are often simple, so performing multiple queries requires the download of the full dataset. This is also due to the fact that the majority of databases do not have APIs that allow programming queries. These general downloads are not always available or only offer a partial dataset. In addition to this, many annotations are made using free text strings, which makes integration and contrasting of information difficult. In this context, the availability of the aforementioned ontology would provide the terms for describing the data of the different resources which would facilitate data exchange and interoperability. The interoperability of the datasets would generate a virtual global repository which would enable a powerful exploitation of the large volume of isolated, existing data about enhancers. Such data interoperability should also be rooted on the FAIR principles (Findable, Accessible, Interoperable and Reusable) for data management [183]. Methodological aspects discussed and proposed by the GREEKC consortium [181] for the development of the Gene Regulation Knowledge Commons would be applicable here. This would also contribute to facilitate to keep track of the evolution of the information about enhancers.

5. Conclusions

There is an increasing interest in the exploitation of information about enhancers for generating new knowledge about regulatory processes due to their potential relation with disorders. We have analyzed the landscape of databases that contain information about enhancers. Our study shows that the resources are highly heterogeneous in the types of information about enhancers, which makes the integrated exploitation of the resources very difficult. The annotation of the data should also be improved to reflect the content of the literature. The development of knowledge models about enhancers and their integration in existing ontologies should contribute to the interoperability of the databases and to improve the usability and the landscape of biological databases with information about enhancer sequences.

Ethical approval

Ethics approval was not required for this study.

CRediT authorship contribution statement

Juan Mulero Hernández: Conceptualization, Methodology, Investigation, Writing - original draft. **Jesualdo Tomás Fernández-Breis:** Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is part of the grants TIN2017-85949-C2-1-R and PID2020-113723RB-C22 funded by MCIN/AEI/10.13039/501100011033/. Juan Mulero is funded by the Ministerio de Ciencia, Innovación y Universidades (FPU18/03264). This work has also been possible thanks to the COST Action CA15205: GREEKC, supported by COST (European Cooperation in Science and Technology).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2022.05.045.

References

[1] Benoist C, Chambon P. In vivo sequence requirements of the sv40 early promoter region. Nature 1981;290(5804):304–10.

[2] Banerji J, Rusconi S, Schaffner W. Expression of a β-globin gene is enhanced by remote sv40 dna sequences. Cell 1981;27(2):299–308.

[3] Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet 2013;14(4):288–95.

[4] Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nature Struct Mol Biol 2014;21(3):210–9.

[5] Maurya SS. Role of enhancers in development and diseases. Epigenomes 2021;5(4):21.

[6] Kvon EZ, Waymack R, Gad M, Wunderlich Z. Enhancer redundancy in development and disease. Nat Rev Genet 2021;22(5):324–36.

[7] Tang F, Yang Z, Tan Y, Li Y. Super-enhancer function and its application in cancer targeted therapy. NPJ Precision Oncol 2020;4(1):1–7.

[8] Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. Trends Mol Med 2021;27(11):1060–73.

[9] van Arensbergen J, van Steensel B, Bussemaker HJ. In search of the determinants of enhancer–promoter interaction specificity. Trends Cell Biol 2014;24(11):695–702.

[10] Chen H, Xiao J, Shao T, Wang L, Bai J, Lin X, Ding N, Qu Y, Tian Y, Chen X, et al. Landscape of enhancer-enhancer cooperative regulation during human cardiac commitment. Mol Therapy-Nucleic Acids 2019;17:840–51.

[11] Barolo S. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. Bioessays 2012;34(2):135–41.

[12] Halfon MS. Studying transcriptional enhancers: the founder fallacy, validation creep, and other biases. Trends Genet 2019;35(2):93–103.

[13] Sabarís G, Laiker I, Noon EP-B, Frankel N. Actors with multiple roles: pleiotropic enhancers and the paradigm of enhancer modularity. Trends Genet 2019;35(6):423–33.

[14] Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop extrusion. Cell Rep 2016;15(9):2038–49.

[15] Ganji M, Shaltiel IA, Bisht S, Kim E, Kalichava A, Haering CH, Dekker C. Real-time imaging of dna loop extrusion by condensin. Science 2018;360(6384):102–5.

[16] Lim B, Levine MS. Enhancer-promoter communication: hubs or loops? Current Opinion Genetics Dev 2021;67:5–9.

[17] Beagrie RA, Pombo A. Gene activation by metazoan enhancers: diverse mechanisms stimulate distinct steps of transcription. Bioessays 2016;38(9):881–93.

[18] Kulaeva OI, Nizovtseva EV, Polikanov YS, Ulianov SV, Studitsky VM. Distant activation of transcription: mechanisms of enhancer action. Mol Cellular Biol 2012;32(24):4892–7.

[19] Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biol 2021;22(1):1–30.

[20] Furlong EE, Levine M. Developmental enhancers and chromosome topology. Science 2018;361(6409):1341–5.

[21] Kolovos P, Knoch TA, Grosveld FG, Cook PR, Papantonis A. Enhancers and silencers: an integrated and simple model for their function. Epigenetics Chromatin 2012;5(1):1–8.

[22] Vernimmen D, Bickmore WA. The hierarchy of transcriptional activation: from enhancer to promoter. Trends Genet 2015;31(12):696–708.

[23] Andersson R. Promoter or enhancer, what's the difference? deconstruction of established distinctions and presentation of a unifying model. Bioessays 2015;37(3):314–23.

[24] Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet 2020;21(5):292–310.

[25] Coppola CJ, Ramaker RC, Mendenhall EM. Identification and function of enhancers in the human genome. Hum Mol Genet 2016;25(R2):R190–7.

[26] Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell 2013;49(5):825–37.

[27] Lam MT, Li W, Rosenfeld MG, Glass CK. Enhancer rnas and regulated transcriptional programs. Trends Biochem Sci 2014;39(4):170–82.

[28] Tippens ND, Vihervaara A, Lis JT. Enhancer transcription: what, where, when, and why? Genes Dev 2018;32(1):1–3.

[29] Arnold PR, Wells AD, Li XC. Diversity and emerging roles of enhancer rna in regulation of gene expression and cell fate. Front Cell Dev Biol 2020;7:377.

[30] Darrow EM, Chadwick BP. Boosting transcription by transcription: enhancer-associated transcripts. Chromosome Res 2013;21(6):713–24.

[31] Han J, Zhang Z, Wang K. 3c and 3c-based techniques: the powerful tools for spatial genome organization deciphering. Mol Cytogenetics 2018;11(1):1–10.

[32] Noordermeer D, Duboule D. Chromatin looping and organization at developmentally regulated gene loci. Wiley Interdisciplinary Rev: Dev Biol 2013;2(5):615–30.

[33] Abbas A, He X, Niu J, Zhou B, Zhu G, Ma T, Song J, Gao J, Zhang MQ, Zeng J. Integrating hi-c and fish data for modeling of the 3d organization of chromosomes. Nature Commun 2019;10(1):1–14.

[34] Brown JM, Roberts NA, Graham B, Waithe D, Lagerholm C, Telenius JM, De Ornellas S, Oudelaar AM, Scott C, Szczerbal I, et al. A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. Nature Commun 2018;9(1):1–15.

[35] Field A, Adelman K. Evaluating enhancer function and transcription. Annu Rev Biochem 2020;89:213–34.

[36] Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. Bmc Genomics 2019;20(1):1–22.

[37] Moore JE, Pratt HE, Purcaro MJ, Weng Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol 2020;21(1):1–16.

[38] Catarino RR, Stark A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev 2018;32(3–4):202–23.
[39] Gargis AS, Kalman L, Bick DP, Da Silva C, Dimmock DP, Funke BH, Gowrisankar S, Hegde MR, Kulkarni S, Mason CE, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nature Biotechnol 2015;33(7):689–93.
[40] Mathelier A, Shi W, Wasserman WW. Identification of altered cis-regulatory elements in human disease. Trends Genet 2015;31(2):67–76.
[41] Y. Murakawa, M. Yoshihara, H. Kawaji, M. Nishikawa, H. Zayed, H. Suzuki, Y. Hayashizaki, F. Consortium, et al., Enhanced identification of transcriptional enhancers provides mechanistic insights into diseases, Trends in Genetics 32 (2) (2016) 76–88..
[42] Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A phase separation model for transcriptional control. Cell 2017;169(1):13–23.
[43] Hong J-W, Hendrix DA, Levine MS. Shadow enhancers as a source of evolutionary novelty. Science 2008;321(5894). 1314–1314.
[44] Cannavò E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, Ciglar L, Korbel JO, Furlong EE. Shadow enhancers are pervasive features of developmental regulatory networks. Curr Biol 2016;26(1):38–51.
[45] Fukaya T, Lim B, Levine M. Enhancer control of transcriptional bursting. Cell 2016;166(2):358–68.
[46] Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 2010;466(7305):490–3.
[47] Waymack R, Fletcher A, Enciso G, Wunderlich Z. Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic. Elife 2020;9:e59351.
[48] Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. Defining functional dna elements in the human genome. Proc Nat Acad Sci 2014;111(17):6131–8.
[49] Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D, et al. Intragenic enhancers act as alternative promoters. Mol Cell 2012;45(4):447–58.
[50] Jiang W, Chen L. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing, Computational and Structural. Biotechnol. J. 2021;19:183–95.
[51] Ott CJ, Suszko M, Blackledge NP, Wright JE, Crawford GE, Harris A. A complex intronic enhancer regulates expression of the cftr gene by direct interaction with the promoter. J Cellular Mol Med 2009;13(4):680–92.
[52] Ahituv N. Exonic enhancers: proceed with caution in exome and genome sequencing studies. Genome Med 2016;8(1):1–3.
[53] Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. An atlas of active enhancers across human cell types and tissues. Nature 2014;507(7493):455–61.
[54] Li J, Hsu A, Hua Y, Wang G, Cheng L, Ochiai H, Yamamoto T, Pertsinidis A. Single-gene imaging links genome topology, promoter–enhancer communication and transcription control. Nature Struct Mol Biol 2020;27 (11):1032–40.
[55] Pott S, Lieb JD. What are super-enhancers? Nature Gen 2015;47(1):8–12.
[56] Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013;153(2):307–19.
[57] Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. Cell 2013;155 (4):934–47.
[58] Quevedo M, Meert L, Dekker MR, Dekkers DH, Brandsma JH, van den Berg DL, Ozgür Z, van IJcken WF, et al. Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. Nature Commun 2019;10(1):1–15.
[59] Peng Y, Zhang Y. Enhancer and super-enhancer: Positive regulators in gene transcription. Animal Models Exp Med 2018;1(3):169–79.
[60] Xiao S, Huang Q, Ren H, Yang M. The mechanism and function of super enhancer rna. Genesis 2021;59(5–6):e23422.
[61] Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, Xu J, Yuan G-C. Dissecting super-enhancer hierarchy based on chromatin interactions. Nature Commun 2018;9(1):1–12.
[62] Raisner R, Bainer R, Haverty PM, Benedetti KL, Gascoigne KE. Super-enhancer acquisition drives oncogene expression in triple negative breast cancer. PLoS One 2020;15(6):e0235343.
[63] Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Mol Cell 2015;58(2):362–70.
[64] Niederriter AR, Varshney A, Parker SC, Martin DM. Super enhancers in cancers, complex disease, and developmental disorders. Genes 2015;6 (4):1183–200.
[65] Jiang Y, Jiang Y-Y, Lin D-C. Super-enhancer-mediated core regulatory circuitry in human cancer, Computational and Structural. Biotechnol J 2021;19:2790–5.
[66] Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, Martchenko A, So V, Macpherson NN, Mitchell JA. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res 2017;27 (2):246–58.
[67] Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science 2014;346(6215):1373–7.
[68] Khan A, Mathelier A, Zhang X. Super-enhancers are transcriptionally more active and cell type-specific than stretch enhancers. Epigenetics 2018;13 (9):910–22.
[69] Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, van Bueren KL, Chines PS, Narisu N, Black BL, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Nat Acad Sci 2013;110(44):17921–6.
[70] Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 2012;148(1–2):84–98.
[71] Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res 2012;22 (3):490–503.
[72] Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. Upstream anti-sense promoters are hubs of transcription factor binding and active histone modifications. Mol Cell 2015;58(6):1101.
[73] Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. Local regulation of gene expression by lncrna promoters, transcription and splicing. Nature 2016;539(7629):452–5.
[74] Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nature Methods 2017;14(6):629–35.
[75] Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJ, Gifford DK, Sherwood RI. High-throughput mapping of regulatory dna. Nature Biotechnol 2016;34(2):167–74.
[76] Dao LT, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, Charbonnier G, Griffon A, Vanhille L, Stephen T, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nature Genetics 2017;49(7):1073–81.
[77] Medina-Rivera A, Santiago-Algarra D, Puthier D, Spicuglia S. Widespread enhancer activity from core promoters. Trends Biochem Sci 2018;43 (6):452–68.
[78] Grosveld F, van Staalduinen J, Stadhouders R. Transcriptional regulation by (super) enhancers: from discovery to mechanisms. Ann Rev Genomics Human Genetics 2021;22:127–46.
[79] Gurumurthy A, Shen Y, Gunn EM, Bungert J. Phase separation and transcription regulation: are super-enhancers and locus control regions primary sites of transcription complex assembly? Bioessays 2019;41(1):1800164.
[80] Gisselbrecht SS, Palagi A, Kurland JV, Rogers JM, Ozadam H, Zhan Y, Dekker J, Bulyk ML. Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. Mol Cell 2020;77 (2):324–37.
[81] Segert JA, Gisselbrecht SS, Bulyk ML. Transcriptional silencers: Driving gene expression with the brakes on. Trends Genet 2021;37(6):514–27.
[82] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473(7345):43–9.
[83] Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res 2011;21 (8):1273–83.
[84] Spicuglia S, Vanhille L. Chromatin signatures of active enhancers. Nucleus 2012;3(2):126–31.
[85] Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, Curina A, Prosperini E, Ghisletti S, Natoli G. Latent enhancers activated by stimulation in differentiated cells. Cell 2013;152(1–2):157–71.
[86] Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods 2012;9(5):473–6.
[87] Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of encode segmentation predictions. Genome Res 2014;24 (10):1595–602.
[88] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. Proc Nat Acad Sci 2010;107(50):21931–6.
[89] Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 2011;470(7333):279–83.
[90] Cruz-Molina S, Respuela P, Tebartz C, Kolovos P, Nikolic M, Fueyo R, van Ijcken WF, Grosveld F, Frommolt P, Bazzi H, et al. Prc2 facilitates the regulatory topology required for poised enhancer function during pluripotent stem cell differentiation. Cell Stem Cell 2017;20(5):689–705.
[91] Crispatzu G, Rehimi R, Pachano T, Bleckwehl T, Cruz-Molina S, Xiao C, Mahabir E, Bazzi H, Rada-Iglesias A. The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. Nature Commun 2021;12(1):1–17.
[92] Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EE. Enhancer loops appear stable during development and are associated with paused polymerase. Nature 2014;512(7512):96–100.
[93] Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature 2012;489(7414):109–13.
[94] Nguyen ML, Jones SA, Prier JE, Russ BE. Transcriptional enhancers in the regulation of t cell differentiation. Front Immunol 2015;6:462.

[95] Libbrecht MW, Rodriguez OL, Weng Z, Bilmes JA, Hoffman MM, Noble WS. A unified encyclopedia of human functional dna elements through fully automated annotation of 164 human cell types. Genome Biol 2019;20(1):1–14.
[96] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518(7539):317–30.
[97] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. Integrative annotation of chromatin elements from encode data. Nucleic Acids Res 2013;41(2):827–41.
[98] Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol 2015;16(1):1–8.
[99] Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet 2020;21(2):71–87.
[100] Lewis MW, Li S, Franco HL. Transcriptional control by enhancers and enhancer rnas. Transcription 2019;10(4–5):171–86.
[101] Xiong L, Kang R, Ding R, Kang W, Zhang Y, Liu W, Huang Q, Meng J, Guo Z. Genome-wide identification and characterization of enhancers across 10 human tissues. Int J Biological Sci 2018;14(10):1321.
[102] Kouno T, Moody J, Kwon AT-J, Shibayama Y, Kato S, Huang Y, Böttcher M, Motakis E, Mendez M, Severin J, et al. C1 cage detects transcription start sites and enhancer activity at single-cell resolution. Nature Commun 2019;10(1):1–12.
[103] Sartorelli V, Lauberth SM. Enhancer rnas are an important regulatory layer of the epigenome. Nature Struct Mol Biol 2020;27(6):521–8.
[104] Melamed P, Yosefzon Y, Rudnizky S, Pnueli L. Transcriptional enhancers: Transcription, function and flexibility. Transcription 2016;7(1):26–31.
[105] Kim T-K, Shiekhattar R. Architectural and functional commonalities between enhancers and promoters. Cell 2015;162(5):948–59.
[106] Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X, et al. Functional roles of enhancer rnas for oestrogen-dependent transcriptional activation. Nature 2013;498(7455):516–20.
[107] Schaukowitch K, Joo J-Y, Liu X, Watts JK, Martinez C, Kim T-K. Enhancer rna facilitates nelf release from immediate early genes. Molecular cell 2014;56(1):29–42.
[108] Sigova AA, Abraham BJ, Ji X, Molinie B, Hannett NM, Guo YE, Jangi M, Giallourakis CC, Sharp PA, Young RA. Transcription factor trapping by rna in gene regulatory elements. Science 2015;350(6263):978–81.
[109] Salamon I, Serio S, Bianco S, Pagiatakis C, Crasto S, Chiariello AM, Conte M, Cattaneo P, Fiorillo L, Felicetta A, et al. Divergent transcription of the nkx2-5 locus generates two enhancer rnas with opposing functions. Iscience 2020;23(9):101539.
[110] Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, Gagneur J. Accurate promoter and enhancer identification in 127 encode and roadmap epigenomics cell types and tissues by genostan. PloS one 2017;12(1):e0169249.
[111] Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA. Integrating diverse datasets improves developmental enhancer prediction. PLoS Comput Biol 2014;10(6):e1003677.
[112] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with chromhmm. Nature Protocols 2017;12(12):2478–92.
[113] Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature 2006;444(7118):499–502.
[114] X. Tang, J. Wang, H. Tao, L. Yuan, G. Du, Y. Ding, K. Xu, X. Bai, Y. Li, Y. Sun, et al., Regulatory patterns analysis of transcription factor binding site clustered regions and identification of key genes in endometrial cancer, Computational and Structural Biotechnology Journal..
[115] Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. Chip-seq accurately predicts tissue-specific activity of enhancers. Nature 2009;457(7231):854–8.
[116] Klein DC, Hainer SJ. Genomic methods in profiling dna accessibility and factor localization. Chromosome Res 2020;28(1):69–85.
[117] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature 2012;489(7414):75–82.
[118] Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen C-A, Lin S, Lin Y, Qiu Y, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 2015;518(7539):350–4.
[119] Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends Genet 2015;31(8):426–33.
[120] Koch F, Andrau J-C. Initiating rna polymerase ii and tips as hallmarks of enhancer activity and tissue-specificity. Transcription 2011;2(6):263–8.
[121] Bae S, Lesch BJ. H3k4me1 distribution predicts transcription state and poising at promoters. Front Cell Dev Biol 2020;8:289.
[122] Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, Imbert J, Andrau J-C, Ferrier P, Spicuglia S. H3k4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J 2011;30(20):4198–210.
[123] Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. Widespread transcriptional pausing and elongation control at enhancers. Genes Dev 2018;32(1):26–41.
[124] Dorighi KM, Swigut T, Henriques T, Bhanu NV, Scruggs BS, Nady N, Still II CD, Garcia BA, Adelman K, Wysocka J. Mll3 and mll4 facilitate enhancer rna synthesis and transcription from promoters independently of h3k4 monomethylation. Mol Cell 2017;66(4):568–76.
[125] Rickels R, Herz H-M, Sze CC, Cao K, Morgan MA, Collings CK, Gause M, Takahashi Y-H, Wang L, Rendleman EJ, et al. Histone h3k4 monomethylation catalyzed by trr and mammalian compass-like proteins at enhancers is dispensable for development and viability. Nature Genetics 2017;49(11):1647–53.
[126] Zhang T, Zhang Z, Dong Q, Xiong J, Zhu B. Histone h3k27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. Genome Biol 2020;21(1):1–7.
[127] Zhao Y, Liu Q, Acharya P, Stengel KR, Sheng Q, Zhou X, Kwak H, Fischer MA, Bradner JE, Strickland SA, et al. High-resolution mapping of rna polymerases identifies mechanisms of sensitivity and resistance to bet inhibitors in t (8; 21) aml. Cell Rep 2016;16(7):2003–16.
[128] Wang J, Zhao Y, Zhou X, Hiebert SW, Liu Q, Shyr Y. Nascent rna sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics 2018;19(1):1–18.
[129] Lopes R, Agami R, Korkmaz G. Gro-seq, a tool for identification of transcripts regulating gene expression. In: Promoter Associated RNA. Springer; 2017. p. 45–55.
[130] Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. Briefings Bioinformatics 2016;17(6):967–79.
[131] Iuchi H, Matsutani T, Yamada K, Iwano N, Sumi S, Hosoda S, Zhao S, Fukunaga T, Hamada M. Representation learning applications in biological sequence analysis. Computational and Structural. Biotechnol J 2021;19:3198–208.
[132] D. Santiago-Algarra, L.T. Dao, L. Pradel, A. España, S. Spicuglia, Recent advances in high-throughput approaches to dissect enhancer function, F1000Research 6..
[133] Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics 2015;106(3):159–64.
[134] Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res 2017;27(1):38–52.
[135] Sethi A, Gu M, Gumusgoz E, Chan L, Yan K-K, Rozowsky J, Barozzi I, Afzal V, Akiyama JA, Plajzer-Frick I, et al. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nature Methods 2020;17(8):807–14.
[136] Diao Y, Li B, Meng Z, Jung I, Lee AY, Dixon J, Maliskova L, Guan K-L, Shen Y, Ren B. A new class of temporarily phenotypic enhancers identified by crispr/cas9-mediated genetic screening. Genome Res 2016;26(3):397–405.
[137] Wu X, Kriz AJ, Sharp PA. Target specificity of the crispr-cas9 system. Quantitative Biol 2014;2(2):59–70.
[138] Sander JD, Joung JK. Crispr-cas systems for editing, regulating and targeting genomes. Nature Biotechnol 2014;32(4):347–55.
[139] Simeonov DR, Gowen BG, Boontanrart M, Roth TL, Gagnon JD, Mumbach MR, Satpathy AT, Lee Y, Bray NL, Chan AY, et al. Discovery of stimulation-responsive immune enhancers with crispr activation. Nature 2017;549(7670):111–5.
[140] Manghwar H, Lindsey K, Zhang X, Jin S. Crispr/cas system: recent advances and future prospects for genome editing. Trends Plant Sci 2019;24(12):1102–25.
[141] Hariprakash JM, Ferrari F. Computational biology solutions to identify enhancers-target gene pairs. Comput Struct Biotechnol. J 2019;17:821–31.
[142] Kyrchanova O, Georgiev P. Mechanisms of enhancer-promoter interactions in higher eukaryotes. Int J Mol Sci 2021;22(2):671.
[143] Suzuki HI, Young RA, Sharp PA. Super-enhancer-mediated rna processing revealed by integrative microrna network analysis. Cell 2017;168(6):1000–14.
[144] F. Tang, Y. Zhang, Q.-Q. Huang, M.-M. Qian, Z.-X. Li, Y.-J. Li, B.-P. Li, Z.-L. Qiu, J.-J. Yue, Z.-Y. Guo, Genome-wide identification and analysis of enhancer-regulated micrornas across 31 human cancers, Frontiers in genetics (2020) 644..
[145] Xu H, Zhang S, Yi X, Plewczynski D, Li MJ. Exploring 3d chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer-promoter interaction. Comput Struct Biotechnol J 2020;18:558–70.
[146] De Wit E, De Laat W. A decade of 3c technologies: insights into nuclear organization. Genes Dev 2012;26(1):11–24.
[147] Giorgetti L, Heard E. Closing the loop: 3c versus dna fish. Genome Biol 2016;17(1):1–9.
[148] Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nat Rev Genet 2019;20(8):437–55.
[149] Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. Mol Cell 2018;71(5):858–71.
[150] Ko JY, Oh S, Yoo KH. Functional enhancers as master regulators of tissue-specific gene regulation and cancer development. Mol Cells 2017;40(3):169.
[151] Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. Nature 2020;583(7818):699–710.
[152] Khan A, Zhang X. dbsuper: a database of super-enhancers in mouse and human genome. Nucleic Acids Res 2016;44(D1):D164–71.
[153] Bai X, Shi S, Ai B, Jiang Y, Liu Y, Han X, Xu M, Pan Q, Wang F, Wang Q, et al. Endb: a manually curated database of experimentally supported enhancers for human and mouse. Nucleic Acids Res 2020;48(D1):D51–7.

[154] Chen C, Zhou D, Gu Y, Wang C, Zhang M, Lin X, Xing J, Wang H, Zhang Y. Sea version 3.0: a comprehensive extension and update of the super-enhancer archive. Nucleic Acids Res 2020;48(D1):D198–203.

[155] Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, Han X, Shi S, Zhang J, Li X, et al. Sedb: a comprehensive human super-enhancer database. Nucleic Acids Res 2019;47(D1):D235–43.

[156] Qian F-C, Li X-C, Guo J-C, Zhao J-M, Li Y-Y, Tang Z-D, Zhou L-W, Zhang J, Bai X-F, Jiang Y, et al. Seanalysis: a web tool for super-enhancer associated regulatory analysis. Nucleic Acids Res 2019;47(W1):W248–55.

[157] Kang R, Tan Z, Lang M, Jin L, Zhang Y, Zhang Y, Guo T, Guo Z. Enhffl: A database of enhancer mediated feed-forward loops for human and mouse. Precision Clinical Med 2021;4(2):129–35.

[158] Cai Z, Cui Y, Tan Z, Zhang G, Tan Z, Zhang X, Peng Y. Raedb: a database of enhancers identified by high-throughput reporter assays. Database 2019.

[159] Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. Database 2017.

[160] Gao T, Qian J. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res 2020;48(D1):D58–64.

[161] Wang Z, Zhang Q, Zhang W, Lin J-R, Cai Y, Mitra J, Zhang ZD. Hedd: human enhancer disease database. Nucleic Acids Res 2018;46(D1):D113–20.

[162] Zeng W, Min X, Jiang R. Endisease: a manually curated database for enhancer-disease associations. Database 2019.

[163] Visel A, Minovitsky S, Dubchak I, Pennacchio LA. Vista enhancer browser-a database of tissue-specific human enhancers. Nucleic Acids Res 2007;35 (suppl_1):D88–92.

[164] Huang M, Wang Y, Yang M, Yan J, Yang H, Zhuang W, Xu Y, Koeffler HP, Lin D-C, Chen X. dbindel: a database of enhancer-associated insertion and deletion variants by analysis of h3k27ac chip-seq. Bioinformatics 2020;36 (5):1649–51.

[165] Kang R, Zhang Y, Huang Q, Meng J, Ding R, Chang Y, Xiong L, Guo Z. Enhancerdb: a resource of transcriptional regulation in the context of enhancers. Database 2019.

[166] Kumar R, Lathwal A, Kumar V, Patiyal S, Raghav PK, Raghava GP. Cancerend: a database of cancer associated enhancers. Genomics 2020;112(5):3696–702.

[167] Zhang Z, Hong W, Ruan H, Jing Y, Li S, Liu Y, Wang J, Li W, Diao L, Han L. Hera: an atlas of enhancer rnas across human tissues. Nucleic Acids Res 2021;49 (D1):D932–8.

[168] Hait TA, Amar D, Shamir R, Elkon R. Focs: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. Genome Biol 2018;19(1):1–14.

[169] Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. Hacer: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res 2019;47 (D1):D106–12.

[170] Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, Wang F, Su J, Wu Q, Liu H, et al. Sea: a super-enhancer archive. Nucleic Acids Res 2016;44(D1):D172–9.

[171] Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MT, Cheng C, Fan X, Gerstein M, et al. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature Genetics 2017;49(10):1428–36.

[172] Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal R, Lupien M, Markowitz S, Scacheri PC, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res 2014;24(1):1–13.

[173] Yang H, Wang S, Xia X. ienhancer-rd: Identification of enhancers and their strength using rkpk features and deep neural networks. Anal Biochem 2021;630:114318.

[174] Yang R, Wu F, Zhang C, Zhang L. ienhancer-gan: a deep learning framework in combination with word embedding and sequence generative adversarial net to identify enhancers and their strength. Int J Mol Sci 2021;22(7):3589.

[175] Li Q, Xu L, Li Q, Zhang L. Identification and classification of enhancers using dimension reduction technique and recurrent neural network. Comput Math Methods Med 2020.

[176] Kamran H, Tahir M, Tayara H, Chong KT. ienhancer-deep: A computational predictor for enhancer sites and their strength using deep learning. Appl Sci 2022;12(4):2120.

[177] Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. Standardized description of scientific evidence using the evidence ontology (eco). Database 2014.

[178] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. Minimum information about a microarray experiment (miame)-toward standards for microarray data. Nature Genetics 2001;29(4):365–71.

[179] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nature Genetics 2000;25(1):25–9.

[180] Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The sequence ontology: a tool for the unification of genome annotations. Genome Biol 2005;6(5):1–12.

[181] Kuiper M, Bonello J, Fernández-Breis JT, Bucher P, Futschik ME, Gaudet P, Kulakovskiy IV, Licata L, Logie C, Lovering RC, et al. The gene regulation knowledge commons: the action area of greekc. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 2022;1865(1):194768.

[182] Sant DW, Sinclair M, Mungall CJ, Schulz S, Zerbino D, Lovering RC, Logie C, Eilbeck K. Sequence ontology terminology for gene regulation. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 2021;1864(10):194745.

[183] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. The fair guiding principles for scientific data management and stewardship. Sci Data 2016;3:1–9.