Review Article

iDarwin volume 6, pages 1-36 Published on April 21, AS0026 (2026 AD)

Integrative, Functional and Artificial Intelligence Frameworks for Discovery of Gene Regulatory Elements and Noncoding Disease Variants

Nilmini HETTIARACHCHI

University of Colombo, Faculty of Science, Thurstan Road, Colombo 03, Sri Lanka

Running title: Functional Genomics of Regulatory Elements and AI

Corresponding author contact: nilminihett@gmail.com

Abstract

The majority of functional variation in the human genome resides within noncoding regions that regulate gene expression through complex cis-regulatory architectures. Among these elements, enhancers play a central role in coordinating spatial and temporal transcriptional programs across development, physiology, and disease. Although enhancers are evolutionarily conserved and functionally indispensable, their systematic identification and characterization have historically been challenging due to their

context-dependent activity and distal genomic locations. Advances in high-throughput functional genomics have transformed enhancer discovery from low-throughput experimental validation to genome-scale quantitative interrogation.
Massively parallel reporter assays (MPRA) and self-transcribing active regulatory region sequencing (STARR-seq) have emerged as powerful experimental platforms for functional enhancer discovery and regulatory variant interpretation. MPRA enables targeted, high-resolution analysis of candidate regulatory sequences and variant effects, whereas STARR-seq allows unbiased genome-wide identification of enhancer activity. Together, these complementary approaches enable large-scale mapping of functional regulatory landscapes across diverse cellular contexts.
The growing availability of large-scale functional genomics datasets has further enabled the application of artificial intelligence and deep learning approaches for predicting enhancer activity directly from DNA sequence and prioritizing disease-associated regulatory variants. In this review, we integrate advances in enhancer biology, high-throughput functional assays, and computational modeling to outline emerging frameworks for discovery and functional interpretation of gene regulatory elements and noncoding disease variants.

Keywords: Gene regulatory elements, enhancers, MPRA, STARR-seq, Deep learning

Introduction

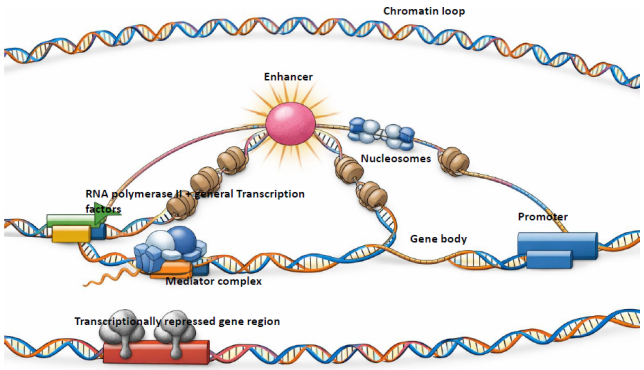
Completion of the human genome sequence marked a major milestone in biology, providing a comprehensive blueprint of protein-coding genes and genomic organization (International Human Genome Sequencing Consortium, 2001; 2004). However, it soon became evident that protein-coding sequences represent only a small fraction of the genome, while the vast majority consists of noncoding DNA with critical regulatory functions (Birney et al., 2007; ENCODE Project Consortium, 2012). Advances in comparative genomics, epigenomics, and transcriptomics have revealed that gene expression is controlled by a complex network of cis-regulatory elements that coordinate transcriptional programs across developmental stages, cell types, and environmental conditions (Lee et al., 2010; Shlyueva et al., 2014). Among these regulatory elements, enhancers play a central role in modulating transcription by integrating transcription factor binding, chromatin accessibility (Figure 1A), and three-dimensional genome architecture (Hnisz et al., 2013; Schoenfelder and Fraser, 2019).

Figure 1

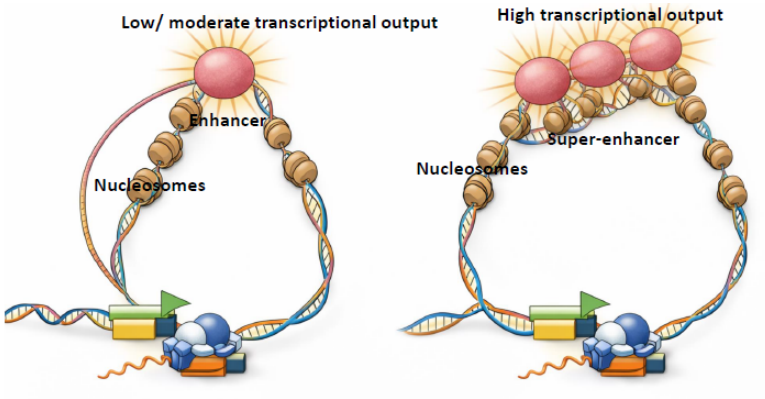
(A) **Regulatory landscape and enhancer–promoter architecture in gene regulation.** Schematic representation of the genomic regulatory landscape illustrating long-range communication between distal enhancers and target promoters through chromatin looping. Enhancers recruit transcription factors and coactivator complexes, including the Mediator complex, which facilitate assembly and stabilization of RNA polymerase II and general transcription factors at promoter regions. Nucleosome positioning and chromatin accessibility influence regulatory activity and transcriptional output. The enhancer–promoter interaction enables precise spatial and temporal control of gene expression. A transcriptionally repressed genomic region is also shown, highlighting the coexistence of active and inactive chromatin states within the regulatory genome.

(B)Enhancers and super-enhancers in transcriptional regulation.
Comparison between a typical enhancer and a super-enhancer in gene regulation. Typical enhancers consist of individual regulatory elements that recruit transcription factors and coactivators to modulate transcription of target genes through chromatin looping interactions with promoters. In contrast, super-enhancers are clusters of closely spaced enhancer elements characterized by dense occupancy of transcription factors, Mediator complexes, and chromatin regulators. These regulatory domains drive high levels of transcriptional activity of genes associated with cell identity, development, and disease. Super-enhancers are frequently enriched at key lineage-defining genes and disease-associated loci, reflecting their central role in controlling gene regulatory networks.

(A)



(B)



Enhancers are typically defined as DNA sequences capable of increasing transcription independent of distance and orientation relative to target promoters (Figure 1B) (Banerji et al., 1981; Shlyueva et al., 2014). They can act over long genomic distances and frequently function in a cell-type–specific manner, enabling precise spatiotemporal regulation of gene expression (Whyte et al., 2013; Andersson and Sandelin, 2020). Disruption of enhancer activity can lead to altered gene expression programs and has been implicated in numerous human diseases, including cancer, developmental disorders, and complex polygenic traits (Sur and Taipale, 2016; Nasser et al., 2021). Genome-wide association studies (GWAS) have further demonstrated that the majority of disease-associated variants reside within noncoding regions, many of which overlap putative enhancer elements (Maurano et al., 2012; Gallagher and Chen-Plotkin, 2018). These findings underscore the importance of systematic identification and functional characterization of regulatory elements for understanding genome function and disease mechanisms. A substantial proportion of genetic variants associated with complex traits and human diseases reside within noncoding regions of the genome, highlighting the central importance of regulatory genomics in understanding genotype–phenotype relationships. Many of these variants that map to candidate regulatory elements such as enhancers, promoters, and insulators, suggest that altered gene regulation represents a key mechanism underlying complex disease susceptibility. Consequently, systematic discovery and functional characterization of regulatory elements has become a primary objective of modern functional and computational genomics.

Despite their importance, functional identification of enhancers has been challenging. Early approaches relied on low-throughput reporter assays and transgenic models, limiting the scale at which regulatory elements could be experimentally validated (Lettice et al., 2003; Pennacchio et al., 2006). While epigenomic profiling techniques such as chromatin accessibility mapping and histone modification analysis have enabled genome-wide identification of candidate regulatory elements, these methods provide indirect evidence of function and cannot definitively establish causal regulatory activity (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium, 2015). Additionally, extensive computational methods and pipelines have been developed for initial identification of the gene regulatory elements which guides the experimental verification (Hettiarachchi 2021; Hettiarachchi and Saitou 2016; Hettiarachchi et al., 2014). Consequently, there has been a growing need for experimental platforms capable of directly measuring enhancer activity across large numbers of sequences.

The development of high-throughput functional reporter assays has transformed enhancer discovery and regulatory genomics. Massively parallel reporter assays (MPRA) enable simultaneous testing of thousands to millions of candidate regulatory

sequences by linking each sequence to unique molecular barcodes and quantifying transcriptional output through high-throughput sequencing (Melnikov et al., 2012; Patwardhan et al., 2012). In parallel, self-transcribing active regulatory region sequencing (STARR-seq) enables genome-wide discovery of enhancers by allowing candidate sequences to transcribe themselves when active (Arnold et al., 2013). These technologies provide quantitative, scalable frameworks for functional interrogation of regulatory elements and have become central tools for interpreting noncoding genomic variation.

More recently, the integration of artificial intelligence and deep learning approaches with functional genomics datasets has further accelerated progress in enhancer discovery. Computational models trained on MPRA and STARR-seq data can predict enhancer activity, identify functional regulatory variants, and infer sequence features governing regulatory grammar (Zhou and Troyanskaya, 2015; Kelley et al., 2016). Transformer-based architectures now model long-range regulatory interactions at megabase scales (Kelley, 2020; Avsec et al., 2021). This convergence of experimental and computational methodologies is enabling increasingly precise mapping of regulatory architecture and facilitating translation of genomic data into biological and clinical insight.

In this review, we provide a comprehensive overview of enhancer biology and examine the principles and applications of MPRA and STARR-seq technologies in functional genomics. We discuss emerging roles of machine learning in regulatory element discovery, highlight applications in disease and precision medicine, and outline key challenges and future directions in the field. By integrating experimental and computational perspectives, this review aims to provide a unified framework for understanding functional enhancer discovery in the post-genomic era.

1.1 Regulatory elements and enhancer biology

Cis-regulatory elements (CREs) orchestrate spatial and temporal gene expression through coordinated interactions between DNA sequence, transcription factors, and chromatin architecture. These regulatory elements include promoters, enhancers, silencers, and insulators, which collectively control transcriptional output across diverse cellular contexts (Levine et al., 2014; Shlyueva et al., 2014). Promoters are typically located proximal to transcription start sites and recruit basal transcriptional machinery, whereas

enhancers can function at variable distances and orientations to increase transcriptional activity (Banerji et al., 1981; Andersson and Sandelin, 2020).

Enhancers represent one of the most extensively studied classes of regulatory elements due to their central role in modulating gene expression during development, differentiation, and disease. Active enhancers are generally characterized by open chromatin accessibility and enrichment of histone modifications such as H3K27ac and H3K4me1 (Heintzman et al., 2009; Creyghton et al., 2010). Genome-wide mapping of these epigenomic signatures has enabled identification of millions of candidate regulatory elements across human cell types (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium, 2015). However, biochemical signatures alone cannot definitively establish functional enhancer activity, highlighting the need for direct experimental validation (Shlyueva et al., 2014).

Enhancers typically function through combinatorial transcription factor binding and recruitment of coactivator complexes such as Mediator and RNA polymerase II. These interactions facilitate transcriptional activation through chromatin remodeling and formation of enhancer–promoter loops within three-dimensional nuclear space (Bulger and Groudine, 2011; Schoenfelder and Fraser, 2019). Chromatin architectural proteins including CTCF and cohesin contribute to the establishment of topologically associating domains (TADs) that constrain enhancer–promoter interactions and influence regulatory specificity (Dixon et al., 2012; Nora et al., 2012). The spatial organization of chromatin therefore plays a critical role in determining enhancer function and target gene selection.

Enhancers frequently operate in clusters rather than as isolated elements. Large clusters of regulatory elements termed super-enhancers exhibit exceptionally high occupancy of transcription factors, Mediator, and chromatin regulators and are strongly associated with genes controlling cell identity (Whyte et al., 2013; Hnisz et al., 2013). Super-enhancers have been implicated in numerous diseases, particularly cancer, where aberrant activation of oncogenic super-enhancers can drive pathological transcriptional programs (Sur and Taipale, 2016). The cooperative and redundant nature of enhancer clusters underscores the complexity of regulatory architecture and the challenges associated with identifying individual functional elements.

A defining feature of enhancer activity is its context dependence. Enhancer function varies across cell types, developmental stages, and environmental conditions due to differences in transcription factor expression and chromatin accessibility (Spitz and Furlong, 2012; Andersson and Sandelin, 2020). The same DNA sequence may act as an enhancer in one cellular context but remain inactive

or even function as a silencer in another. This context specificity complicates prediction of enhancer activity based solely on sequence or chromatin features and highlights the importance of functional assays capable of measuring regulatory activity directly.

Comparative genomics has also provided important insights into enhancer evolution and function. Many enhancers are evolutionarily conserved across vertebrates, reflecting functional constraints and essential regulatory roles (Pennacchio et al., 2006; Villar et al., 2015). At the same time, lineage-specific enhancers contribute to species-specific traits and regulatory innovation (Reilly et al., 2015). Integration of evolutionary conservation with functional assays has therefore become an important strategy for identifying regulatory elements with biological significance.

1.2 Regulatory evolution and enhancer diversification

The evolution of gene regulatory elements has played a central role in shaping phenotypic diversity across species. Changes in cis-regulatory sequences, particularly enhancers, are increasingly recognized as major drivers of evolutionary innovation because they can modulate gene expression without altering protein-coding sequences (Carroll, 2008; Wray, 2007). Comparative genomic studies have revealed that while many protein-coding genes are highly conserved across species, regulatory elements exhibit substantial evolutionary turnover, reflecting dynamic processes of enhancer gain, loss, and modification.

Enhancer evolution occurs through multiple mechanisms, including sequence mutation, transcription factor binding site turnover, transposable element insertion, and duplication of regulatory modules. Even small sequence changes can alter transcription factor binding affinity and modify enhancer strength, leading to quantitative differences in gene expression that may be subject to natural selection (Prud'homme et al., 2007; Villar et al., 2015). Such regulatory variation provides a flexible substrate for evolutionary adaptation while minimizing deleterious pleiotropic effects often associated with coding mutations.

Comparative functional genomics studies have demonstrated both conservation and divergence of enhancer activity across species. Many developmental enhancers show deep evolutionary conservation, reflecting strong functional constraints on regulatory programs controlling fundamental biological processes (Pennacchio et al., 2006). At the same time, lineage-specific enhancers contribute to species-specific traits and regulatory innovation. Genome-wide enhancer mapping across vertebrates has

revealed that enhancer repertoires evolve rapidly, with frequent turnover of regulatory elements even when target gene expression patterns remain conserved (Villar et al., 2015).

Transposable elements represent another important source of regulatory innovation. Many enhancers originate from transposable element sequences that acquire regulatory function following insertion into new genomic contexts (Chuong et al., 2017). Co-option of transposable elements into regulatory networks can generate novel transcription factor binding sites and contribute to species-specific gene expression patterns. This process highlights the dynamic and modular nature of enhancer evolution.

The integration of evolutionary analysis with functional genomics is essential for understanding the regulatory basis of phenotypic diversity and disease susceptibility. Regulatory variants that influence enhancer activity may be subject to evolutionary selection, and some disease-associated variants may reflect evolutionary trade-offs between adaptive and deleterious effects (Wray, 2007). Functional assays capable of measuring enhancer activity across orthologous sequences and variant alleles provide powerful tools for dissecting these evolutionary dynamics.

As high-throughput functional assays and comparative genomics datasets continue to expand, the study of regulatory evolution is transitioning from descriptive analysis to quantitative functional investigation. Integrating evolutionary perspectives with experimental and computational approaches will enhance understanding of how regulatory elements evolve, how they shape gene expression networks, and how regulatory variation contributes to both adaptive traits and human disease.

While epigenomic mapping and computational prediction have generated extensive catalogs of candidate enhancers, functional validation remains essential for distinguishing active regulatory elements from nonfunctional sequences. Functional interrogation of regulatory evolution has been greatly accelerated by high-throughput enhancer assays, which now enable quantitative testing of regulatory sequence divergence across species. High-throughput functional assays such as Massively Parallel Reporter Assays (MPRA) and STARR-seq provide scalable frameworks for quantitative evaluation of enhancer activity and have transformed the study of gene regulation at genome-wide scale. By comparing orthologous sequences across species, these approaches allow quantification of functional divergence in regulatory activity and identification of sequence features underlying evolutionary changes. Massively parallel reporter assays have been used to evaluate the functional impact of human-specific substitutions within conserved regulatory elements, revealing how small sequence changes can alter transcriptional output

(Kheradpour et al., 2013). Similarly, comparative STARR-seq analyses across species have provided insights into the conservation and turnover of enhancer activity within regulatory landscapes (Arnold et al., 2014).

.

2. Massively parallel reporter assays (MPRA)

Massively parallel reporter assays (MPRAs) have emerged as a transformative technology for functional interrogation of cis-regulatory elements at scale. Traditional reporter assays enabled experimental validation of individual enhancers but were limited by low throughput and labor-intensive experimental design. The development of MPRA enabled simultaneous functional testing of thousands to millions of candidate regulatory sequences within a single experiment, thereby transforming enhancer discovery into a quantitative and scalable endeavor (Melnikov et al., 2012; Patwardhan et al., 2012).

The conceptual framework of MPRA is based on linking candidate regulatory sequences to unique molecular barcodes within reporter constructs. These barcodes allow precise quantification of transcriptional output associated with each sequence through high-throughput sequencing of RNA relative to DNA input. Early implementations demonstrated that synthetic and genomic enhancer sequences could be systematically tested for regulatory activity in parallel, enabling quantitative dissection of enhancer strength and regulatory architecture (Melnikov et al., 2012). Subsequent studies extended this approach to in vivo mammalian systems and demonstrated the feasibility of large-scale enhancer functional screening (Patwardhan et al., 2012).

A typical MPRA workflow involves the design and synthesis of candidate regulatory sequences, association with unique barcode sequences, cloning into reporter vectors, delivery into relevant cell types, and sequencing-based quantification of reporter expression. Enhancer activity is estimated as the ratio of barcode counts in RNA relative to plasmid DNA, providing a quantitative measure of regulatory output (Inoue and Ahituv, 2015; Tewhey et al., 2016). Advances in array-based DNA synthesis have enabled construction of highly complex libraries containing hundreds of thousands of sequences, facilitating systematic interrogation of regulatory elements across the genome (Figure 2).

One of the most powerful applications of MPRA is the functional characterization of noncoding variants associated with human disease. Genome-wide association studies have revealed that most trait-associated variants reside in noncoding regions, often within putative enhancers. MPRA enables direct comparison of reference and alternative alleles to quantify their effects on transcriptional activity, allowing identification of causal regulatory variants within associated loci (Tewhey et al., 2016; Ulirsch et

al., 2016). This approach has been widely used to interpret variants linked to immune disorders, hematological traits, and metabolic phenotypes.

MPRA has also enabled detailed dissection of regulatory grammar through systematic mutagenesis and motif perturbation. Saturation mutagenesis MPRA experiments introduce single-nucleotide substitutions across enhancer sequences to map functional transcription factor binding sites and cooperative interactions (Patwardhan et al., 2009; Kheradpour et al., 2013). Such analyses have revealed that enhancer activity is often governed by combinatorial motif interactions, spacing constraints, and local sequence context. These findings contribute to understanding of how regulatory information is encoded within DNA sequence.

Variants of MPRA have expanded its utility across diverse biological systems. Self-transcribing MPRA designs eliminate the need for external barcodes, while viral delivery systems enable stable genomic integration of reporter constructs. More recently, single-cell MPRA approaches have combined reporter assays with single-cell transcriptomics, allowing measurement of enhancer activity across heterogeneous cellular populations (Muerdter et al., 2018; Klein et al., 2020). These innovations are enabling increasingly refined analysis of context-dependent regulatory activity.

Despite its strengths, MPRA has several limitations. Most implementations rely on episomal reporter constructs that may not fully recapitulate endogenous chromatin context or long-range genomic interactions. Enhancer activity can also be influenced by promoter choice within reporter constructs, potentially affecting measured transcriptional output (Inoue et al., 2017). Additionally, synthetic oligonucleotide constraints may limit tested sequence length, potentially excluding distal regulatory interactions. These limitations highlight the importance of integrating MPRA results with complementary approaches such as CRISPR-based perturbation and chromatin profiling.

Nevertheless, MPRA remains a cornerstone technology for functional genomics and enhancer discovery. By enabling systematic and quantitative interrogation of regulatory sequences, MPRA bridges the gap between genomic annotation and functional validation. The large-scale datasets generated by MPRA experiments have also become critical resources for training computational models that predict enhancer activity and variant effects, further expanding its impact across regulatory genomics.

3. STARR-seq: genome-wide functional enhancer discovery

Self-transcribing active regulatory region sequencing (STARR-seq) has emerged as one of the most powerful experimental approaches for unbiased genome-wide identification of enhancers. While massively parallel reporter assays (MPRA) enable targeted functional testing of candidate regulatory sequences, STARR-seq allows direct screening of genomic fragments for enhancer activity without prior selection, thereby enabling comprehensive mapping of functional regulatory elements (Arnold et al., 2013).

STARR-seq was first developed in Drosophila as a high-throughput reporter assay capable of identifying enhancers across entire genomes. In this approach, candidate genomic fragments are cloned downstream of a minimal promoter within a reporter construct. If a fragment possesses enhancer activity, it drives transcription from the promoter and is transcribed as part of the reporter RNA. The enhancer sequence therefore functions as its own molecular barcode, allowing direct quantification of activity through sequencing of reporter transcripts (Arnold et al., 2013). This self-transcribing design distinguishes STARR-seq from conventional reporter assays and enables quantitative measurement of enhancer activity across millions of sequences simultaneously.

Subsequent adaptations of STARR-seq have enabled its application in mammalian systems, including human and mouse cells, revealing extensive regulatory landscapes across diverse cell types (Arnold et al., 2014; Muerdter et al., 2018). Genome-wide STARR-seq studies have identified numerous enhancers lacking canonical chromatin signatures, highlighting the importance of functional assays for accurate regulatory annotation (Arnold et al., 2013; Barakat et al., 2018). These findings demonstrate that biochemical markers such as histone modifications and chromatin accessibility alone may not fully capture functional enhancer activity.

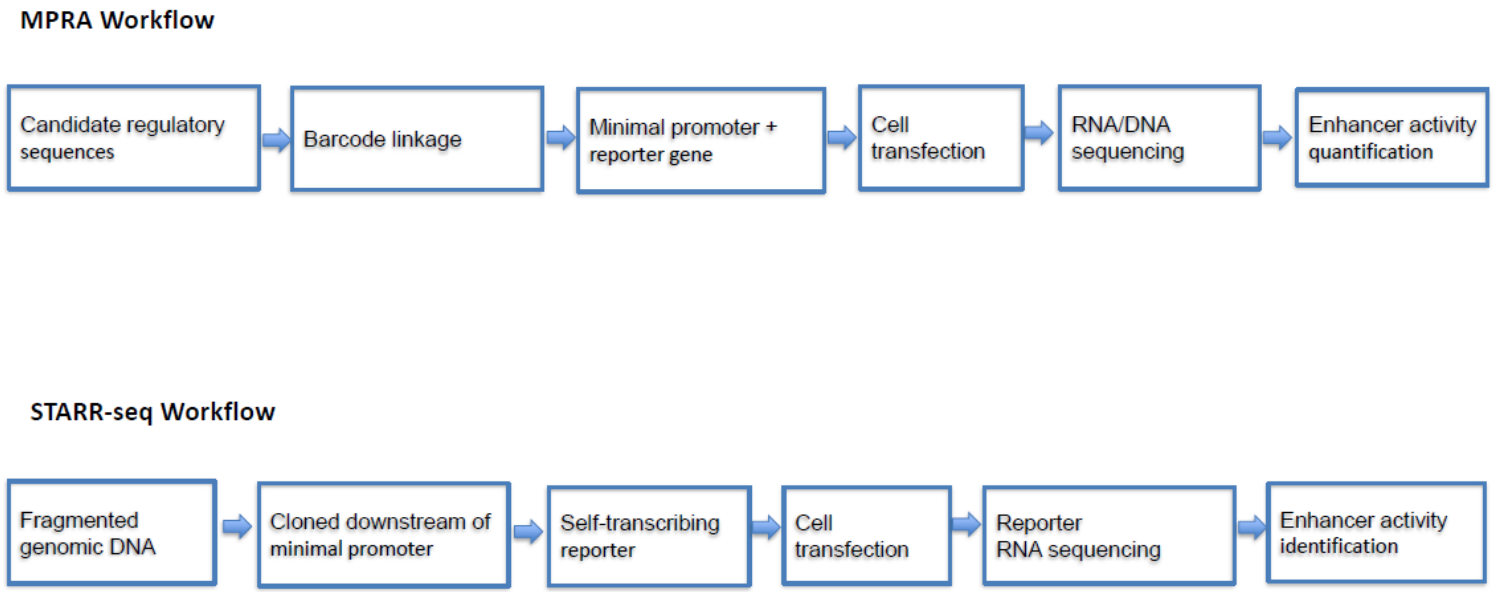
As illustrated in Figure 2 STARR-seq workflow typically involves fragmentation of genomic DNA, cloning into reporter vectors, transfection into relevant cell types, RNA and DNA extraction, and sequencing-based quantification of enhancer activity. Enhancer strength is estimated by comparing RNA abundance of each fragment relative to plasmid DNA input (Arnold et al., 2013; Liu et al., 2017). Improvements in library preparation, sequencing depth, and computational analysis have enhanced sensitivity and reproducibility of STARR-seq experiments.

Several methodological refinements have expanded the utility of STARR-seq. Targeted STARR-seq approaches focus on specific genomic regions of interest, enabling deeper coverage and improved sensitivity for disease-associated loci (Vockley et al.,

2016). Incorporation of unique molecular identifiers and improved normalization methods has reduced amplification bias and enhanced quantitative accuracy (Muerdter et al., 2018). More recently, adaptations integrating STARR-seq with single-cell transcriptomics are beginning to enable measurement of enhancer activity at single-cell resolution, offering insights into regulatory heterogeneity across cell populations.

Figure 2. Experimental design and workflow of MPRA and STARR-seq

Experimental design and workflow of massively parallel reporter assays (MPRA) and STARR-seq. In MPRA, candidate regulatory sequences are linked to unique DNA barcodes upstream of a minimal promoter and reporter gene. Enhancer activity is quantified by sequencing barcode abundance in RNA relative to DNA. In STARR-seq, fragmented genomic DNA is cloned downstream of a minimal promoter within the reporter transcript. Active enhancers drive transcription of themselves, enabling direct quantification through reporter RNA sequencing.



STARR-seq has been widely applied to investigate regulatory mechanisms in development, evolution, and disease. Genome-wide enhancer maps generated by STARR-seq have revealed context-dependent regulatory activity across cell types and environmental conditions (Arnold et al., 2014; Barakat et al., 2018). Comparative analyses across species have provided insights into enhancer evolution and lineage-specific regulatory innovations. In cancer genomics, STARR-seq has been used to identify tumor-specific enhancers and regulatory networks driving oncogene expression.

Despite its strengths, STARR-seq shares limitations with other episomal reporter assays. Because genomic fragments are tested outside their endogenous chromatin context, measured activity may not fully reflect native enhancer function. Fragment size and promoter choice can also influence detection sensitivity and regulatory output (Muerdter et al., 2018). Additionally, STARR-seq may exhibit reduced sensitivity for weak enhancers or elements requiring specific chromosomal context for activity. Integration with complementary approaches such as MPRA, CRISPR-based perturbation, and chromatin conformation mapping is therefore essential for comprehensive characterization of regulatory elements.

Nevertheless, STARR-seq remains one of the most comprehensive tools for genome-wide functional enhancer discovery. Its ability to identify regulatory elements independent of prior annotation has revealed a more complex and dynamic regulatory landscape than previously appreciated. When combined with targeted assays and computational modeling, STARR-seq provides critical insights into the architecture and function of cis-regulatory elements across genomes.

4. Comparative evaluation of MPRA, STARR-seq and integrative functional genomics approaches

Massively parallel reporter assays (MPRA) and self-transcribing active regulatory region sequencing (STARR-seq) represent complementary experimental frameworks for large-scale enhancer discovery. While both technologies quantify transcriptional output driven by candidate regulatory sequences, they differ substantially in experimental design, scope, and optimal applications (Melnikov et al., 2012; Arnold et al., 2013). A comparative evaluation of these approaches is essential for selecting appropriate strategies in regulatory genomics.

A primary distinction between MPRA and STARR-seq lies in their approach to candidate sequence selection. MPRA is typically a targeted, hypothesis-driven assay in which candidate enhancers are selected based on prior annotation such as chromatin accessibility, histone modifications, evolutionary conservation, or disease association (Inoue and Ahituv, 2015; Tewhey

et al., 2016). This design enables systematic interrogation of specific regulatory regions and precise testing of sequence variants. In contrast, STARR-seq allows unbiased genome-wide screening by directly cloning fragmented genomic DNA, enabling functional discovery of enhancers without prior annotation (Arnold et al., 2013; Liu et al., 2017). This makes STARR-seq particularly valuable for comprehensive enhancer mapping in defined cellular contexts.

MPRA offers greater flexibility for dissecting regulatory grammar at high resolution. Synthetic oligonucleotide libraries enable saturation mutagenesis, motif perturbation, and combinatorial testing of regulatory elements, allowing detailed mapping of transcription factor binding sites and sequence constraints (Patwardhan et al., 2009; Kheradpour et al., 2013). By contrast, STARR-seq typically evaluates longer genomic fragments, which are advantageous for capturing broader regulatory context but may limit fine-scale mutational analysis unless combined with targeted designs.

Both technologies provide quantitative measurements of enhancer strength, but their experimental configurations introduce distinct considerations. MPRA relies on barcode-based quantification of transcriptional output, whereas STARR-seq uses the enhancer sequence itself as a transcriptional barcode (Arnold et al., 2013; Melnikov et al., 2012). Comparative analyses have demonstrated that experimental design choices—including promoter selection, vector configuration, and library complexity—can substantially influence measured enhancer activity (Muerdter et al., 2018; Klein et al., 2020). These findings highlight the importance of standardized design and careful interpretation of reporter assay data.

In the context of disease genomics, MPRA has become particularly powerful for systematic interpretation of noncoding variants. Parallel testing of reference and alternative alleles allows direct quantification of variant effects on transcriptional activity, enabling prioritization of causal variants within genome-wide association study loci (Tewhey et al., 2016; Ulirsch et al., 2016). STARR-seq has also contributed to variant interpretation, particularly in targeted implementations; however, MPRA remains more widely used for allele-specific dissection due to its flexible synthetic design.

A shared limitation of both MPRA and STARR-seq is their reliance on episomal reporter constructs in most implementations. Comparative analyses have shown substantial differences between chromosomal and episomal enhancer activity measurements, underscoring the influence of genomic context (Inoue et al., 2017). Endogenous chromatin architecture, three-dimensional genome organization, and promoter compatibility can all influence enhancer function, motivating integration with CRISPR-based perturbation assays and chromatin conformation capture techniques (Schoenfelder and Fraser, 2019).

The convergence of these experimental approaches with computational modeling further enhances their impact. Deep learning models trained on MPRA and STARR-seq datasets enable prediction of enhancer activity and variant effects across the genome (Zhou and Troyanskaya, 2015; Avsec et al., 2021). Computational predictions can prioritize candidate sequences for functional validation, while high-throughput assays provide empirical data to refine model accuracy. This iterative experimental–computational cycle represents a central paradigm in modern regulatory genomics.

Rather than viewing MPRA and STARR-seq as competing technologies, it is more productive to consider them as complementary components of an integrated functional genomics framework. STARR-seq excels at unbiased genome-wide enhancer discovery, while MPRA provides high-resolution mechanistic dissection and variant testing. Integration with CRISPR perturbation, three-dimensional genome mapping, and AI-driven prediction enables increasingly comprehensive characterization of regulatory elements. Together, these approaches are reshaping our understanding of enhancer architecture and function (Table 1).

Table 1. Comparison of High-Throughput Enhancer Assays

Feature	Classical Reporter Assay	MPRA	STARR-seq
Throughput	Low (1–10 sequences)	10^3–10^6	10^6–10^8 fragments
Barcode strategy	None	External DNA barcode	Self-transcribing insert
Sequence selection	Manual	Targeted/synthetic	Genome-wide (unbiased)
Quantitative output	Semi-quantitative	RNA/DNA ratio	RNA/DNA ratio

Feature	Classical Reporter Assay	MPRA	STARR-seq
Variant testing	Limited	Excellent (allelic, saturation mutagenesis)	Possible but less flexible
Episomal vs genomic	Episomal	Episomal (mostly)	Episomal
Best use case	Mechanistic validation	Variant interpretation, regulatory grammar	Genome-wide enhancer discovery
Limitations	Low scale	Requires library design	Fragment context loss

5. Applications in disease biology and functional genomics

5.1 Enhancer dysregulation as a driver of human disease

A growing body of evidence indicates that dysregulation of enhancer activity plays a central role in human disease. Genome-wide association studies (GWAS) have revealed that the majority of disease-associated variants reside within noncoding regions of the genome, frequently overlapping candidate regulatory elements such as enhancers (Maurano et al., 2012; Gallagher and Chen-Plotkin, 2018). These findings have shifted the focus of human genetics from protein-coding mutations toward regulatory variation as a major determinant of complex traits and disease susceptibility (Albert and Kruglyak, 2015; Zhang and Lupski, 2015).

Enhancers influence gene expression in a cell-type–specific and context-dependent manner, and even modest alterations in enhancer activity can lead to measurable phenotypic consequences. Unlike coding mutations that directly alter protein structure, regulatory variants often exert quantitative effects on gene expression, contributing to disease risk through subtle transcriptional

changes (Nasser et al., 2021). High-throughput functional assays such as MPRA and STARR-seq have therefore become essential tools for translating statistical genetic associations into mechanistic understanding.

5.2 Functional interpretation of noncoding variants

MPRA has emerged as a powerful platform for systematic functional evaluation of noncoding variants identified through GWAS. By testing reference and alternative alleles in parallel reporter constructs, MPRA enables direct measurement of variant effects on enhancer activity and gene regulation (Tewhey et al., 2016; Ulirsch et al., 2016). This approach has been widely applied to prioritize candidate causal variants within disease-associated loci and to elucidate regulatory mechanisms underlying complex traits.

Functional testing of disease-associated variants using MPRA has provided insights into immune disorders, hematological traits, metabolic disease, and cardiovascular risk (Tewhey et al., 2016; Ulirsch et al., 2016). Integration of MPRA data with expression quantitative trait locus (eQTL) mapping and chromatin accessibility profiles strengthens causal inference and enables identification of regulatory variants that modulate transcription factor binding or enhancer strength (Farh et al., 2015). These integrative analyses are critical for understanding how noncoding variation contributes to disease susceptibility.

STARR-seq has also contributed to interpretation of disease-associated regulatory variation by enabling genome-wide mapping of active enhancers in disease-relevant cell types. Targeted STARR-seq approaches focusing on GWAS loci have identified functional enhancer elements associated with disease risk and gene expression changes (Vockley et al., 2016; Liu et al., 2017). Although less commonly used for systematic allele-by-allele testing than MPRA, STARR-seq provides valuable context for identifying regulatory hotspots within disease-associated regions.

5.3 Cancer and enhancer reprogramming

Cancer genomes exhibit widespread regulatory reprogramming, including activation of oncogenic enhancers and super-enhancers that drive aberrant transcriptional programs. Super-enhancers associated with oncogenes such as MYC and lineage-specific

transcription factors can promote tumor progression and therapeutic resistance (Hnisz et al., 2013; Sur and Taipale, 2016). Functional characterization of these regulatory regions is therefore critical for understanding tumor biology.

STARR-seq has been applied to identify cancer-specific enhancers and regulatory networks active in tumor cells but absent in normal counterparts (Barakat et al., 2018). MPRA-based mutational scanning has further enabled fine-scale dissection of oncogenic enhancer elements, revealing sequence features required for malignant transcriptional activity. Such studies demonstrate that enhancer dysregulation can function as a primary driver of oncogenesis.

The therapeutic potential of targeting enhancer-mediated transcription has also gained attention. Pharmacological inhibition of transcriptional coactivators associated with super-enhancers has shown promise in preclinical cancer models, highlighting the translational relevance of enhancer biology (Whyte et al., 2013; Bradner et al., 2017).

5.4 Developmental and neurological disorders

Enhancers play a crucial role in controlling gene expression during development, and mutations in regulatory elements can lead to congenital disorders and developmental abnormalities. Functional enhancer assays have been used to identify regulatory elements controlling key developmental genes and to characterize variants associated with developmental syndromes (Pennacchio et al., 2006; Nasser et al., 2021).

In neurological and neuropsychiatric disorders, many disease-associated variants map to noncoding regions active in neuronal cell types. MPRA studies in neural progenitor cells and differentiated neurons have identified regulatory variants affecting synaptic genes and transcriptional regulators involved in brain development and function (Schaub et al., 2012; Forrest et al., 2014). Integration of functional enhancer data with chromatin interaction maps is particularly important in neurological systems, where long-range regulatory interactions are common.

5.5 Precision medicine and therapeutic applications

Functional interpretation of regulatory variation has important implications for precision medicine. Identifying causal enhancer variants can improve risk prediction, inform therapeutic targeting, and guide development of regulatory-element–based

biomarkers. High-throughput functional assays also enable engineering of synthetic enhancers with predictable activity profiles for gene therapy and biotechnology applications (Inoue and Ahituv, 2015).

In synthetic biology, MPRA-guided design of regulatory elements has been used to construct promoters and enhancers with tunable expression levels. Integration of deep learning models with experimental validation further enables rational design of regulatory sequences with desired functional properties (Brookes et al., 2022). These advances highlight the translational potential of enhancer discovery technologies beyond basic research.

5.6 Integrative functional genomics pipelines

The most powerful applications of MPRA and STARR-seq arise from integrative workflows combining experimental and computational approaches. A typical pipeline may involve identification of disease-associated loci through GWAS, annotation of candidate enhancers using epigenomic datasets, functional testing using MPRA or STARR-seq, computational prioritization using deep learning models, and validation using CRISPR-based perturbation (Nasser et al., 2021). Such integrative strategies accelerate translation from genomic association to mechanistic understanding.

As functional genomics technologies continue to evolve, the integration of high-throughput reporter assays, genome editing, and computational modeling will play a central role in elucidating regulatory mechanisms underlying human disease and phenotypic diversity (Table 2).

6. Artificial intelligence and deep learning in enhancer discovery

The rapid expansion of functional genomics datasets has catalyzed the integration of artificial intelligence (AI) and deep learning approaches into regulatory genomics. High-throughput functional assays such as MPRA and STARR-seq generate quantitative measurements of enhancer activity across hundreds of thousands of sequences, providing ideal training data for computational models capable of learning regulatory patterns directly from DNA sequence (Zhou and Troyanskaya, 2015; Kelley et al., 2016). These developments have transformed enhancer discovery from descriptive annotation toward predictive modeling of regulatory function.

Table 2. Applications of MPRA and STARR-seq in Disease Genomics

Application Area	MPRA	STARR-seq	Example Insights
GWAS variant prioritization	✓✓✓	✓	Identification of causal SNPs in enhancers
Cancer enhancer reprogramming	✓	✓✓✓	Discovery of tumor-specific enhancers
Developmental biology	✓✓	✓✓	Stage-specific enhancer activation
Synthetic enhancer design	✓✓✓	✓	Engineering regulatory circuits
Regulatory grammar dissection	✓✓✓	✓✓	Motif spacing & combinatorial logic

Early computational methods for enhancer prediction relied primarily on sequence conservation, motif enrichment, and integration of epigenomic features such as chromatin accessibility and histone modifications (Pennacchio et al., 2007; Ernst and Kellis, 2012). While informative, these approaches often lacked the capacity to capture complex combinatorial interactions among transcription factor binding sites and sequence context. The emergence of deep learning architectures has enabled modeling of hierarchical regulatory features without explicit feature engineering, significantly improving predictive performance (Min et al., 2017; Zou et al., 2019).

Convolutional neural network (CNN)-based models such as DeepSEA demonstrated that regulatory activity, transcription factor binding, and chromatin accessibility can be predicted directly from primary DNA sequence (Zhou and Troyanskaya, 2015). Subsequent frameworks including Basset further improved prediction of chromatin accessibility across cell types using deep

convolutional architectures (Kelley et al., 2016). These models provided early evidence that the regulatory code governing enhancer activity is encoded within DNA sequence and can be computationally decoded.

More recent models have incorporated long-range sequence context and advanced neural network architectures. Basenji extended sequence-based prediction to kilobase-scale genomic contexts, enabling modeling of distal regulatory effects on gene expression (Kelley, 2020). Transformer-based models such as Enformer have further improved predictive accuracy by integrating long-range chromatin interactions and attention mechanisms, enabling gene expression prediction across megabase-scale genomic regions (Avsec et al., 2021). These advances represent a major step toward predictive regulatory genomics.

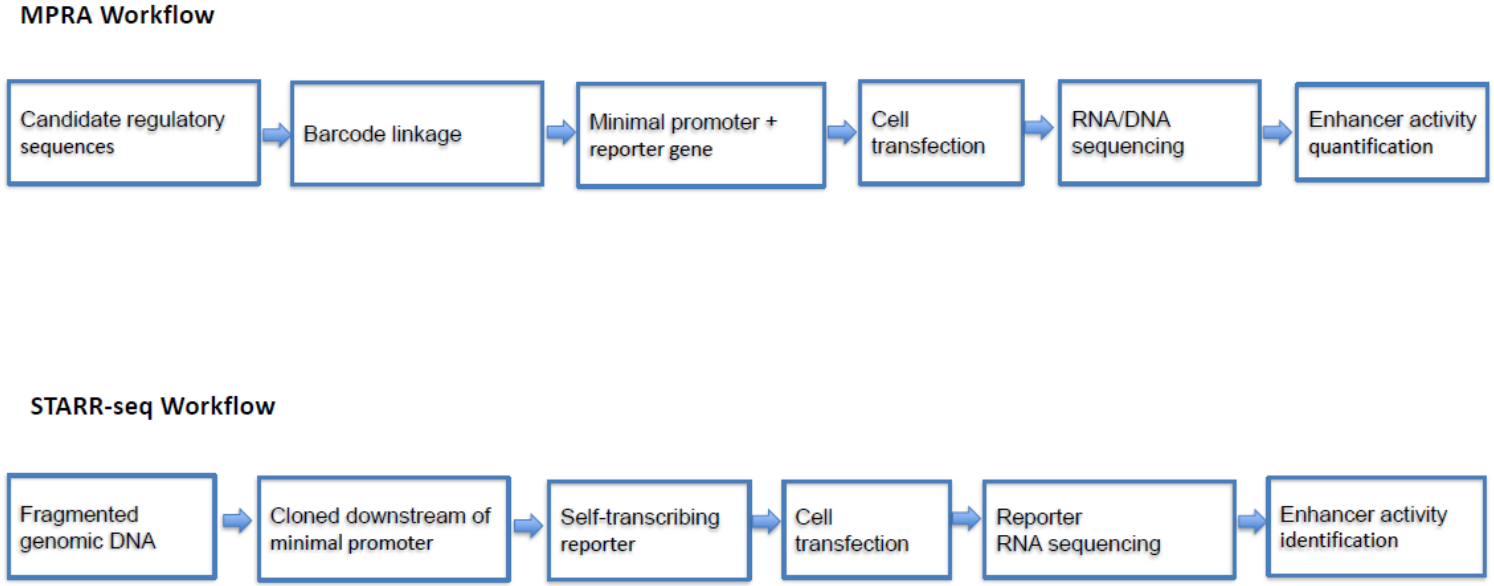
High-throughput functional datasets generated by MPRA and STARR-seq play a critical role in training and benchmarking these models. Unlike purely epigenomic datasets, MPRA and STARR-seq provide direct quantitative measurements of enhancer activity, allowing models to learn causal sequence–function relationships. Deep learning models trained on MPRA data have been used to predict enhancer strength, identify functional transcription factor motifs, and quantify the effects of noncoding variants (Kheradpour et al., 2013; Tewhey et al., 2016). Integration of STARR-seq datasets further improves model generalizability and genome-wide enhancer prediction.

A major application of AI in regulatory genomics is the interpretation of noncoding genetic variation. Genome-wide association studies have identified thousands of disease-associated variants located within noncoding regions, yet functional interpretation remains challenging. Deep learning models can predict how sequence variants alter enhancer activity, transcription factor binding, and chromatin accessibility, enabling prioritization of candidate causal variants (Zhou and Troyanskaya, 2015; Avsec et al., 2021). Coupling computational predictions with MPRA validation has proven particularly effective for identifying functional regulatory variants underlying complex traits.

Beyond prediction, generative AI approaches are beginning to enable design of synthetic regulatory elements. Generative adversarial networks and variational auto encoder frameworks can create novel DNA sequences predicted to exhibit desired enhancer activity profiles (Gupta and Zou, 2019; Brookes et al., 2022). Integration of generative modeling with MPRA-based validation creates an iterative experimental–computational loop for exploring regulatory sequence space and refining predictive models. Integration and application of high-throughput functional assays with AI for enhancer discovery is fully illustrated in Figure 3.

Figure 3. Integration of high-throughput functional assays with AI for enhancer discovery

High-throughput functional datasets from MPRA and STARR-seq are integrated with deep learning frameworks including CNNs (DeepSEA, Basset), transformer-based models (Basenji, Enformer), multi-modal sequence–chromatin integration, and generative models. These approaches enable enhancer activity prediction, regulatory variant interpretation, and synthetic regulatory element design, with iterative experimental validation and model refinement.



Despite these advances, several challenges remain. Deep learning models require large and diverse training datasets and may exhibit limited interpretability. Understanding the sequence features driving model predictions remains an active area of research, with approaches such as in silico mutagenesis and saliency mapping used to identify functional motifs and regulatory grammar (Koo and Eddy, 2019). Additionally, integrating sequence-based predictions with cell-type–specific chromatin context and three-dimensional genome organization remains a key challenge.

Nevertheless, the convergence of AI and high-throughput functional genomics is reshaping enhancer biology. Predictive models trained on MPRA and STARR-seq data are enabling increasingly accurate identification of functional regulatory elements and interpretation of noncoding variation. As datasets grow and models improve, AI-driven regulatory genomics is expected to play a central role in decoding gene regulatory architecture and translating genomic information into biological and clinical insight (Table 3).

7. Challenges and limitations of high-throughput enhancer assays and computational prediction

Despite the transformative impact of high-throughput functional assays and computational modeling on enhancer biology, several conceptual and technical challenges remain. A critical understanding of these limitations is essential for accurate interpretation of enhancer activity and regulatory variant effects.

7.1 Episomal versus endogenous genomic context

One of the primary limitations of both MPRA and STARR-seq is their reliance on episomal reporter constructs in most implementations. In these systems, candidate regulatory sequences are tested outside their native chromosomal environment, lacking endogenous chromatin context and long-range genomic interactions. Comparative analyses have demonstrated substantial differences between chromosomal and episomal measurements of enhancer activity, indicating that reporter assays primarily capture intrinsic regulatory potential rather than full endogenous function (Inoue et al., 2017; Muerdter et al., 2018).

Table 3. Deep Learning Models for Enhancer Prediction

Model	Architecture	Input	Output	Strength
DeepSEA	CNN	DNA sequence	TF binding, chromatin marks	Variant effect prediction
Basset	CNN	DNA	Accessibility	Cell-type modeling
Basenji	CNN (long-range)	DNA (long context)	Expression prediction	Distal enhancer effects
Enformer	Transformer	DNA (200kb+)	Gene expression	Long-range modeling
MPRA-trained models	CNN/Hybrid	MPRA data	Enhancer strength	Causal sequence–function

Enhancer activity in vivo is influenced by chromatin accessibility, DNA methylation, nucleosome positioning, and three-dimensional genome organization. Chromatin looping and topologically associating domains constrain enhancer–promoter interactions and influence target gene specificity (Dixon et al., 2012; Schoenfelder and Fraser, 2019). Because episomal reporter constructs lack this spatial context, measured activity may not fully reflect native regulatory function. Integration with CRISPR-based perturbation methods that assess enhancer necessity within endogenous genomic context is therefore critical for comprehensive functional interpretation (Fulco et al., 2016; Gasperini et al., 2019).

7.2 Promoter dependency and regulatory compatibility

Enhancer activity measured in reporter assays can be influenced by the promoter used in the construct. Many MPRA and STARR-seq experiments utilize minimal promoters to reduce background transcription; however, enhancer–promoter compatibility varies across genomic contexts (Inoue et al., 2017). Certain enhancers exhibit promoter specificity and may function differently depending on promoter architecture. This dependency can introduce bias when comparing enhancer strength across studies or experimental systems.

7.3 Sequence length and genomic context constraints

Synthetic oligonucleotide libraries used in MPRA often restrict tested sequences to relatively short fragments, typically 150–300 base pairs. Many endogenous enhancers, however, extend over larger genomic regions and may contain multiple cooperative regulatory modules (Shlyueva et al., 2014). Testing truncated fragments may fail to capture full enhancer context or long-range interactions. STARR-seq partially addresses this limitation by using longer genomic fragments but may still separate enhancers from neighboring regulatory elements that modulate activity.

7.4 Cell-type specificity and context dependence

Enhancer function is highly context dependent and influenced by transcription factor expression, chromatin accessibility, and cellular state. Reporter assays performed in a single cell type provide only a partial view of regulatory activity. A sequence that functions as a strong enhancer in one cellular context may be inactive in another (Spitz and Furlong, 2012; Andersson and Sandelin, 2020). Comprehensive characterization of enhancer function therefore requires testing across multiple cell types and conditions.

Single-cell functional genomics approaches are beginning to address this limitation by enabling measurement of enhancer activity across heterogeneous cellular populations (Klein et al., 2020). However, these methods remain technically complex and computationally demanding.

7.5 Limitations of deep learning and predictive models

Artificial intelligence and deep learning approaches have demonstrated remarkable success in predicting enhancer activity from DNA sequence, yet several challenges remain. Predictive models are dependent on training data and may exhibit limited generalizability across cell types or experimental conditions (Zou et al., 2019). In addition, deep neural networks often lack interpretability, making it difficult to extract biologically meaningful insights from model predictions (Koo and Eddy, 2019).

Although transformer-based architectures can model long-range regulatory interactions, accurately integrating three-dimensional genome organization and dynamic chromatin context remains challenging (Avsec et al., 2021). Experimental validation using functional assays is therefore essential for confirming computational predictions.

7.6 Data integration and reproducibility

The increasing volume of functional genomics data presents challenges for data integration and reproducibility. Differences in experimental design, sequencing depth, normalization strategies, and computational pipelines can complicate cross-study comparisons (Muerdter et al., 2018). Standardization of experimental protocols and analysis frameworks will be essential for maximizing reproducibility and utility of MPRA and STARR-seq datasets.

7.7 Toward integrated functional genomics

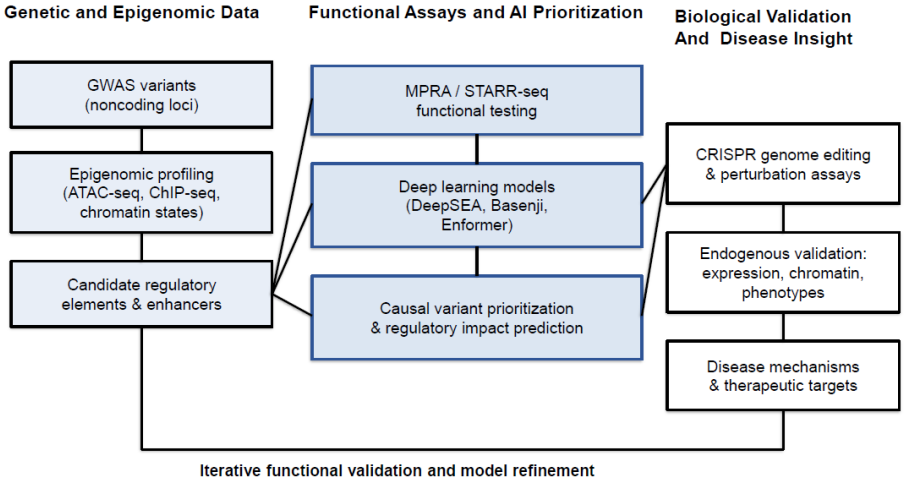
Addressing these challenges will require integrative experimental designs that combine reporter assays with endogenous genome editing, chromatin profiling, and computational modeling. Rather than relying on any single approach, future enhancer discovery efforts will benefit from multi-modal strategies capable of capturing regulatory function across diverse genomic and cellular contexts.

8. Future directions and emerging technologies in enhancer discovery

The field of regulatory genomics is rapidly evolving toward integrative and predictive frameworks for understanding enhancer function. Advances in high-throughput functional assays, genome editing technologies, and artificial intelligence are converging to enable increasingly precise mapping and modeling of gene regulatory elements (Figure 4). Future developments are expected to move beyond static catalogs of enhancers toward dynamic, context-aware models capable of predicting regulatory outcomes across cellular states and environmental conditions.

Figure 4. Integrated experimental and computational framework for functional interpretation of noncoding regulatory variants in disease.

The majority of disease-associated variants identified through genome-wide association studies reside within noncoding regions of the genome, frequently overlapping candidate regulatory elements. Epigenomic profiling identifies putative enhancers and regulatory regions associated with disease loci. High-throughput functional assays such as MPRA and STARR-seq enable quantitative evaluation of enhancer activity and variant effects. Deep learning models trained on functional genomics datasets further prioritize candidate causal variants and predict regulatory impact. Subsequent genome editing and CRISPR-based perturbation assays validate functional effects in endogenous genomic context. This integrative framework enables systematic dissection of regulatory mechanisms underlying complex diseases and accelerates translation of noncoding genomic variation into biological insight.



8.1 Single-cell functional enhancer assays

A major frontier in enhancer discovery is the transition from bulk measurements to single-cell resolution. Conventional MPRA and STARR-seq experiments measure average enhancer activity across heterogeneous cell populations, potentially obscuring cell-type–specific regulatory dynamics. Recent advances integrating reporter assays with single-cell RNA sequencing enable measurement of enhancer activity at single-cell resolution, allowing identification of cell-type–specific regulatory elements and heterogeneity in enhancer usage (Klein et al., 2020; Zhao et al., 2023).

Single-cell functional genomics approaches are particularly valuable for studying developmental processes, immune responses, and tumor heterogeneity, where regulatory programs vary across cellular subpopulations. Integration of single-cell enhancer assays with chromatin accessibility and transcriptomic data will provide a multidimensional view of regulatory landscapes and enable reconstruction of gene regulatory networks across differentiation trajectories (Buenrostro et al., 2015; Cusanovich et al., 2018).

8.2 CRISPR-based endogenous enhancer perturbation

Understanding enhancer necessity requires functional testing within native genomic context. CRISPR-based technologies have emerged as powerful tools for endogenous enhancer interrogation. CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa) enable targeted repression or activation of regulatory elements without altering DNA sequence, while CRISPR-mediated deletions and base editing allow precise modification of enhancer regions (Fulco et al., 2016; Gasperini et al., 2019).

High-throughput CRISPR screens targeting candidate enhancers can be combined with transcriptomic readouts to identify regulatory elements controlling gene expression. Integration of CRISPR perturbation with MPRA and STARR-seq provides complementary evidence of enhancer sufficiency and necessity, enabling comprehensive functional characterization of regulatory elements (Simeonov et al., 2017). Continued development of multiplexed genome editing technologies will further enhance the scalability and precision of endogenous enhancer studies.

8.3 Integration with three-dimensional genome architecture

Enhancer function is tightly linked to three-dimensional genome organization. Chromatin looping and topologically associating domains constrain enhancer–promoter interactions and influence regulatory specificity (Dixon et al., 2012; Schoenfelder and Fraser, 2019). Advances in chromatin conformation capture techniques, including Hi-C and Capture-C, are enabling increasingly detailed mapping of spatial genome organization.

Future enhancer discovery frameworks will integrate functional reporter assays with 3D genome data to predict enhancer–promoter interactions and regulatory outcomes more accurately. Computational models capable of incorporating spatial genome information alongside sequence features will improve understanding of regulatory architecture and the effects of structural genomic variation on gene expression (Dekker et al., 2013; Nasser et al., 2021).

8.3 AI-driven predictive and generative regulatory modeling

Artificial intelligence and deep learning are expected to play an increasingly central role in regulatory genomics. Transformer-based architectures and multi-modal models integrating sequence and epigenomic data can now predict gene expression and regulatory activity with high accuracy (Avsec et al., 2021; Zou et al., 2019). Continued improvements in model interpretability and training data diversity will enhance predictive reliability and biological insight.

Generative AI approaches capable of designing synthetic regulatory elements represent a particularly promising direction. Generative adversarial networks and other deep generative models can create novel DNA sequences predicted to exhibit desired enhancer activity profiles (Gupta and Zou, 2019; Brookes et al., 2022). Coupling these approaches with MPRA-based validation creates an iterative experimental–computational loop for refining regulatory models and exploring sequence space.

Large-scale foundation models trained on genomic and epigenomic data may ultimately provide comprehensive representations of regulatory logic across cell types and species. Such models could predict functional effects of noncoding variants, guide genome editing strategies, and enable rational design of regulatory elements for therapeutic applications.

8.4 Multi-omic integration and evolutionary perspectives

Future enhancer discovery efforts will increasingly rely on integration of multi-omic datasets, including chromatin accessibility, histone modifications, transcription factor occupancy, and transcriptomics. Combining MPRA and STARR-seq data with these complementary datasets will improve identification of functional regulatory elements and clarify context-specific regulatory mechanisms (Roadmap Epigenomics Consortium, 2015).

Comparative and evolutionary genomics will also play an important role in understanding enhancer function. Cross-species comparisons can identify conserved regulatory elements and lineage-specific innovations, providing insights into how regulatory evolution contributes to phenotypic diversity and disease susceptibility (Villar et al., 2015). Integration of functional assays with evolutionary analysis will enhance identification of biologically important regulatory sequences.

8.5 Toward predictive regulatory genomics and clinical translation

The ultimate goal of enhancer discovery is to develop predictive models of gene regulation that can be translated into clinical and biotechnological applications. Functional interpretation of regulatory variants has implications for precision medicine, gene therapy, and synthetic biology. Engineered enhancers with predictable activity profiles could enable tissue-specific expression of therapeutic genes and improved design of gene regulatory circuits (Inoue and Ahituv, 2015).

The convergence of high-throughput functional assays, genome editing technologies, and AI-driven modeling is driving regulatory genomics toward a predictive and translational framework. Continued interdisciplinary collaboration will be essential for realizing the full potential of enhancer biology in medicine and biotechnology.

9 Conclusion

The past two decades have witnessed a paradigm shift in our understanding of genome function, moving from a protein-centric view toward recognition of the central role of noncoding regulatory elements in shaping gene expression, development, and disease. Enhancers, once defined primarily by classical low-throughput reporter assays, are now appreciated as dynamic, context-

dependent regulatory modules embedded within complex chromatin and three-dimensional genomic landscapes. The convergence of high-throughput functional assays and computational modeling has fundamentally transformed enhancer biology from descriptive annotation to quantitative and predictive science.

Massively parallel reporter assays (MPRA) and self-transcribing active regulatory region sequencing (STARR-seq) have emerged as cornerstone technologies in this transformation. MPRA provides a highly flexible and quantitative framework for hypothesis-driven interrogation of regulatory sequences, enabling systematic dissection of enhancer grammar, saturation mutagenesis, and allele-specific variant testing. STARR-seq complements this targeted approach by enabling unbiased, genome-wide identification of functional enhancers directly from genomic DNA. Together, these technologies bridge the gap between epigenomic annotation and causal regulatory function, offering scalable solutions to decode cis-regulatory logic.

Importantly, the large quantitative datasets generated by MPRA and STARR-seq have catalyzed the integration of artificial intelligence and deep learning into regulatory genomics. Sequence-based predictive models trained on functional assay data now approach the capacity to infer enhancer activity directly from DNA sequence. These advances extend beyond prediction; they enable prioritization of noncoding variants, interpretation of genome-wide association study loci, and rational design of synthetic regulatory elements. The iterative feedback loop between experimental validation and computational modeling represents a powerful framework for accelerating discovery.

Applications of these approaches have already reshaped disease genomics. Functional interrogation of noncoding variants has clarified mechanisms underlying cancer, autoimmune disorders, cardiovascular disease, and neuropsychiatric conditions. The recognition that regulatory variation contributes substantially to disease risk underscores the necessity of systematic enhancer characterization. High-throughput reporter assays, integrated with CRISPR-based endogenous perturbation and multi-omic profiling, are enabling a more precise mapping of regulatory architecture underlying complex phenotypes.

Despite these advances, important challenges remain. Episomal assay context, promoter dependency, sequence length constraints, and cell-type specificity complicate interpretation of reporter-based measurements. Deep learning models, while increasingly powerful, require careful validation and improved interpretability. Addressing these limitations will require continued integration of complementary methodologies, including endogenous genome editing, single-cell functional assays, and three-

dimensional genome mapping. Standardization of experimental design and computational pipelines will further enhance reproducibility and cross-study comparability.

Looking forward, the field is poised to transition from cataloging regulatory elements to constructing predictive, context-aware models of gene regulation. The integration of single-cell technologies, high-resolution chromatin conformation mapping, and foundation-scale genomic AI models promises a more comprehensive understanding of enhancer function across development, evolution, and disease. Generative approaches capable of designing synthetic enhancers with specified regulatory properties highlight the translational potential of this work in gene therapy and synthetic biology.

Ultimately, decoding the regulatory genome is central to understanding the molecular basis of organismal complexity and human disease. MPRA, STARR-seq, and AI-driven modeling collectively provide the experimental and computational infrastructure necessary to achieve this goal. Continued interdisciplinary collaboration between molecular biologists, geneticists, computational scientists, and clinicians will be essential for translating regulatory genomics insights into tangible biomedical applications. As the boundaries between experimental and computational biology continue to blur, the systematic functional dissection of enhancers will remain a central endeavor in the post-genomic era.

Conflict of Interest

The author declares no conflict of interest.

Funding

No external funding was received for this work.

Author contributions

The author conceived the review topic, performed the literature survey, synthesized and interpreted the scientific literature, prepared the figures and tables, and wrote and revised the manuscript.

Ethics statement

This manuscript is a review article that synthesizes findings from previously published literature. No new experiments involving human participants, animals, or clinical data were conducted by the author. Consequently, ethical approval and informed consent were not required.

References

Albert Frank W. and Kruglyak Leonid (2015) The role of regulatory variation in complex traits and disease. Nature Reviews Genetics, 16:197–212.

Andersson Robin and Sandelin Albin (2020) Determinants of enhancer and promoter activities of regulatory elements. Nature Reviews Genetics, 21:71–87.

Arnold Christoph D. et al. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science, 339:1074–1077.

Arnold Christoph D. et al. (2014) Quantitative genome-wide enhancer activity maps across species reveal conservation and turnover. Nature Genetics, 46:685–692.

Avsec Žiga et al. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. Nature, 590:290–295.

Banerji Jitendra, Rusconi Stefano, and Schaffner Walter (1981) Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell, 27:299–308.

Barakat Tarik S. et al. (2018) Functional dissection of the enhancer repertoire in human embryonic stem cells. Cell Stem Cell, 23:276–288.

Birney Ewan et al. (2007) Identification and analysis of functional elements in 1% of the human genome. Nature, 447:799–816.

Bradner James E., Hnisz Denes, and Young Richard A (2017) Transcriptional addiction in cancer. Cell, 168:629–643.

Brookes David H. et al. (2022) Generative modeling for designing regulatory DNA. Nature Machine Intelligence, 4:777–787.

Buenrostro Jason D. et al. (2015) Single-cell chromatin accessibility reveals regulatory variation. Nature, 523:486–490.
Bulger Marjorie and Groudine Mark (2011) Functional and mechanistic diversity of distal transcription enhancers. Cell, 144:327–339.
Carroll Sean B. (2008) Evo-devo and an expanding evolutionary synthesis. Cell, 134:25–36.
Chuong Edward B., Elde Nels C., and Feschotte Cedric (2017) Regulatory activities of transposable elements. Nature Reviews Genetics, 18:71–86.
Creyghton Mark P. et al. (2010) Histone H3K27ac separates active from poised enhancers. PNAS, 107:21931–21936.
Cusanovich Darren A. et al. (2018) A single-cell atlas of chromatin accessibility. Science 348:910–914.
Dekker Job et al. (2013) The 4D nucleome project. Nature, 549:219–226.
Dixon Jesse R. et al. (2012) Topological domains in mammalian genomes. Nature, 485:376–380.
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489: 57-74.
Ernst Jason and Kellis Manolis (2012) ChromHMM: automating chromatin-state discovery. Nature Methods, 9:215–216.
Farh Kyle K.-H. et al. (2015) Genetic and epigenetic fine mapping of causal variants in autoimmune disease. Nature, 518:337–343.
Forrest Alistair R. R. et al. (2014) A promoter-level mammalian expression atlas. Nature, 507:462–470.
Fulco Charles P. et al. (2016) Systematic mapping of functional enhancer–promoter connections. Science, 354:769–773.
Gallagher Mark D. and Chen-Plotkin Alice S. (2018) The post-GWAS era: from association to function. American Journal of Human Genetics, 102:717–730.
Gasperini Molly et al. (2019) A genome-wide framework for mapping gene regulation. Cell, 176:377–390.
Gupta Ankit and Zou James (2019) Feedback GAN for DNA sequence design. Nature Machine Intelligence, 1:105–111.
Heintzman Nathaniel D. et al. (2009) Histone modifications at human enhancers reflect global gene expression. Nature, 459:108–112.
Hettiarachchi Nilmini et al. (2014) Lineage specific conserved noncoding sequences of plant genomes. Genome Biol Evol, 9:2527–2542.
Hettiarachchi Nilmini and Saitou Naruya (2016) GC content heterogeneity transition of conserved noncoding sequences. Genome Biol Evol, 8:3377–3392.

Hettiarachchi Nilmini (2021) Primate deep conserved noncoding sequences and non-coding RNA. iDarwin, 2:3–33.
Hnisz Denes et al. (2013) Super-enhancers in the control of cell identity and disease. Cell, 155:934–947.
Inoue Fumitaka and Ahituv Nadav (2015) Decoding enhancers using massively parallel reporter assays. Genomics, 106:159–164.
Inoue Fumitaka et al. (2017) Chromosomal versus episomal encoding of enhancer activity. Genome Research, 27:38–52.
Kelley David R. (2020) Predicting regulatory activity across chromosomes with convolutional neural networks. Nature Methods, 17:107–114.
Kelley David R., Snoek Jasper, and Rinn John L. (2016) Basset: learning the regulatory code. Genome Research, 26:990–999.
Kheradpour Payman et al. (2013) Systematic dissection of regulatory motifs in human enhancers. Genome Research, 23:800–811.
Klein Jason C. et al. (2020) Evaluation of massively parallel reporter assay design. Nature Methods, 17:1083–1091.
Koo Peter K. and Eddy Sean R. (2019) Representation learning of genomic sequence motifs. Genome Biology, 20:114.

Lee Taekjip I. and Young Richard A. (2013) Transcriptional regulation and its misregulation in disease. Cell, 152:1237–1251.
Levine Michael, Cattoglio Claudia, and Tjian Robert (2014) Looping back to leap forward. Cell, 157:13–25.
Lettice Laura A. et al. (2003) A long-range Shh enhancer regulates limb development. Nature, 425:80–84.
Liu Yang et al. (2017) Functional assessment of human enhancer activities using STARR-seq. Genome Biology, 18:219.
Maurano Matthew T. et al. (2012) Localization of disease-associated variation. Science, 337:1190–1195.
Melnikov Alexey et al. (2012) Dissection of inducible enhancers using MPRA. Nature Biotechnology, 30:271–277.
Min Seonwoo, Lee Byunghan, and Yoon Sungroh (2017) Deep learning in bioinformatics. Briefings in Bioinformatics, 18:851–869.
Muerdter Felix et al. (2018) Resolving systematic errors in enhancer assays. Nature Methods, 15:141–149.
Nasser Joseph et al. (2021) Genome-wide enhancer maps link risk variants. Nature, 593:238–243.
Nora Elphège P. et al. (2012) Spatial partitioning of regulatory landscapes. Nature, 485:381–385.
Patwardhan Rajiv P. et al. (2009) High-resolution analysis of DNA regulatory elements. Nature Biotechnology, 27:1173–1175.
Patwardhan Rajiv P. et al. (2012) Functional dissection of mammalian enhancers. Nature Biotechnology, 30:265–270.
Pennacchio Len A. et al. (2006) In vivo enhancer analysis. Nature, 444:499–502.
Prud'Homme Benjamin, Gompel Nicolas, and Carroll Sean B. (2007) Regulatory evolution principles. PNAS, 104:8605–8612.

Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518: 317-330.
Schaub Marc A. et al. (2012) Linking disease associations with regulatory information. Genome Research, 22:1748–1759.
Schoenfelder Stefan and Fraser Peter (2019) Enhancer–promoter contacts. Nature Reviews Genetics, 20:437–455.
Simeonov Dimitre R. et al. (2017) Discovery of stimulation-responsive enhancers. Nature, 549:111–115.
Spitz François and Furlong Eileen E M (2012) Transcription factors in development. Nature Reviews Genetics, 13:613–626.
Sur Inderpreet and Taipale Jussi (2016) Enhancers in cancer. Nature Reviews Cancer, 16:483–493.
Tewhey Ryan et al. (2016) Identification of expression-modulating variants. Cell, 165:1519–1529.
Ulirsch Jonathan C. et al. (2016) Functional dissection of genetic variation. Cell, 165:1530–1545.
Villar Diego et al. (2015) Enhancer evolution across mammals. Cell, 160:554–566.
Vockley Christopher M. et al. (2016) Regulatory effects of noncoding variation. Genome Research, 26:1203–1214.
Whyte Warren A. et al. (2013) Master transcription factors establish super-enhancers. Cell, 153:307–319.
Wray Gregory A. (2007) Evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics, 8:206–216.
Zhao Siqi et al. (2023) A single cell massively parallel reporter assay detects cell type specific gene regulation. Nature Genetics, 55:346-354.
Zhou Jian and Troyanskaya Olga G (2015) Predicting effects of noncoding variants. Nature Methods, 12:931–934.
Zou James et al. (2019) Deep learning in genomics. Nature Genetics, 51:12–18.