REVIEW

Decoding biology with massively parallel reporter assays and machine learning

Alyssa La Fleur,[1] Yongsheng Shi,[2] and Georg Seelig[1,3]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA; [2]Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, Irvine, California 92697, USA; [3]Department of Electrical & Computer Engineering, University of Washington, Seattle, Washington 98195, USA

Massively parallel reporter assays (MPRAs) are powerful tools for quantifying the impacts of sequence variation on gene expression. Reading out molecular phenotypes with sequencing enables interrogating the impact of sequence variation beyond genome scale. Machine learning models integrate and codify information learned from MPRAs and enable generalization by predicting sequences outside the training data set. Models can provide a quantitative understanding of *cis*-regulatory codes controlling gene expression, enable variant stratification, and guide the design of synthetic regulatory elements for applications from synthetic biology to mRNA and gene therapy. This review focuses on *cis*-regulatory MPRAs, particularly those that interrogate cotranscriptional and post-transcriptional processes: alternative splicing, cleavage and polyadenylation, translation, and mRNA decay.

Introduction and historical perspective

A key challenge in the postgenomic era is understanding the relationship between genomic sequence and biological function. In particular, a thorough understanding of how *cis*-regulatory codes govern protein production is critical to linking genetic variation to gene expression changes or designing synthetic regulatory elements for applications from mRNA therapy to synthetic biology. Although significant progress has been made constructing these sequence-to-function links, many challenges remain. Gene expression is a multistep process, and *cis*-regulatory information controlling it is densely encoded, making it difficult to disentangle regulatory codes controlling different processes. Simultaneously, regulatory information controlling a process is often spread across multiple coding and noncoding regions. Moreover, the human genome contains a finite number of genes to learn *cis*-regulatory codes from. Although human population genetic variation can provide additional data, there is a high degree of sequence similarity between individuals, severely limiting sequence representation (Starita et al. 2017).

Massively parallel reporter assays (MPRAs) are a powerful approach for studying gene regulation, overcoming some of the limitations above (Kinney and McCandlish 2019; Trauernicht et al. 2020; Gallego Romero and Lea 2023). In an MPRA, the activity of a biological process of interest is monitored based on reporter expression. A high degree of sequence variation is introduced into this reporter to generate a reporter library. Libraries are delivered into cell extracts, cells, or animals where reporter expression results in a molecular phenotype. Finally, the reporters and their associated molecular phenotypes are quantified by high-throughput sequencing.

The two defining features of an MPRA are that (1) sequence variation is targeted to a region/regions within a reporter gene (e.g., a UTR, intron, or exon) or close to it on the same plasmid or vector (e.g., an enhancer), whereas other features of the reporter construct remain fixed, and (2) the molecular phenotype of interest is read out by sequencing, often using the abundance of a separately encoded barcode as a proxy. Limiting variation to one part of the gene makes it possible to isolate that region's contribution to the process of interest. The parallelism of modern sequencing technologies enables screening of thousands to millions of reporter variants in a single experiment, potentially exceeding the degree of variation found in the genome.

The concepts underlying MPRAs can be traced back to in vitro mutagenesis and selection studies where pools of partially randomized DNA or RNA molecules were synthesized in vitro and subjected to selection based on ligand binding or in vitro biochemical activities (Oliphant and Struhl 1989; Ellington and Szostak 1990; Tuerk and Gold 1990). Given the sequencing capacity limitations of the Sanger sequencing era, it was necessary to select a small number of "winners" for sequencing

[*Keywords*: gene regulation; machine learning; massively parallel reporter assays]
Corresponding author: gseelig@uw.edu, yongshes@uci.edu
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.351800.124. Freely available online through the *Genes & Development* Open Access option.

through multiple rounds of amplification and selection (Ellington and Szostak 1990; Tuerk and Gold 1990). The concept of coupling mutagenesis with selection was later adapted to in-cell reporter assays and found even broader applications (Chen and Chasin 1993; Wang et al. 2004). These methods have significantly contributed to characterizing protein–DNA/RNA binding specificities and identifying regulatory DNA/RNA *cis*-regulatory elements (CREs).

The advent of next-generation sequencing technology unleashed the potential of functional library screening and made MPRAs possible. Early assays targeted sequence variation to promoters (Patwardhan et al. 2009; Kinney et al. 2010; Kwasnieski et al. 2012; Sharon et al. 2012; Mogno et al. 2013; Van Arensbergen et al. 2017), enhancers (Melnikov et al. 2012; Patwardhan et al. 2012; Arnold et al. 2013; Kheradpour et al. 2013), and exons (Ke et al. 2011), aiming to map the influence of *cis*-regulatory sequences on transcription and cassette exon inclusion. Soon after, MPRAs were adapted to characterize variation in protein-coding sequences (Kosuri et al. 2013), 5′ UTRs (Dvir et al. 2013; Noderer et al. 2014; Cuperus et al. 2017; Cambray et al. 2018), and 3′ UTRs (Fig. 1A; Oikonomou et al. 2014; Zhao et al. 2014). Foreshadowing the broad applicability of MPRAs but also the requirement for systems compatible with efficient delivery of large libraries, early experiments were performed in cell-free settings (Patwardhan et al. 2009), the bacterium *Escherichia coli* (Kinney et al. 2010; Kosuri et al. 2013), the yeast *Saccharomyces cerevisiae* (Sharon et al. 2012; Dvir et al. 2013), human cell lines (Melnikov et al. 2012; Oikonomou et al. 2014; Zhao et al. 2014), and mouse retina (White et al. 2013) and through tail vein injection in mice (Fig. 1B; Patwardhan et al. 2012).

MPRA data, with their scale and ability to evenly cover a sequence space of interest, are highly suited for training machine learning models. Breakthroughs in artificial intelligence (AI) and machine learning (ML) applied to natural language processing or image recognition showed that increasing the size and quality of training data sets is as vital to improving performance as model architecture (Hoffmann et al. 2022). MPRA data can similarly drive improvements in the quality of sequence-to-function models in genomics. Machine learning models are powerful tools for learning relationships in the data, allowing generalization beyond sequences characterized in the MPRAs. Even early MPRA publications were sometimes specifically designed to facilitate modeling efforts, highlighting the opportunity that such data sets provide for model training but also underscoring that large-scale data sets can be unwieldy to understand without a quantitative model (Kinney et al. 2010; Dvir et al. 2013; Mogno et al. 2013; Rosenberg et al. 2015). In practice, models and experiments progressed from focusing on individual regulatory elements (Kinney et al. 2010; Melnikov et al. 2012; Patwardhan et al. 2012; Mogno et al. 2013; Noderer et al. 2014) to classical machine learning (Dvir et al. 2013; Rosenberg et al. 2015) and then neural network models (Cuperus et al. 2017; Paggi et al. 2017; Bogard et al. 2019; Movva et al. 2019; Sample et al. 2019; Vainberg Slutskin et al. 2019), aiming to capture generalizable *cis*-regulatory codes.

Here, we review MPRA studies investigating pre-mRNA processing, mRNA stability, and translation (Fig.

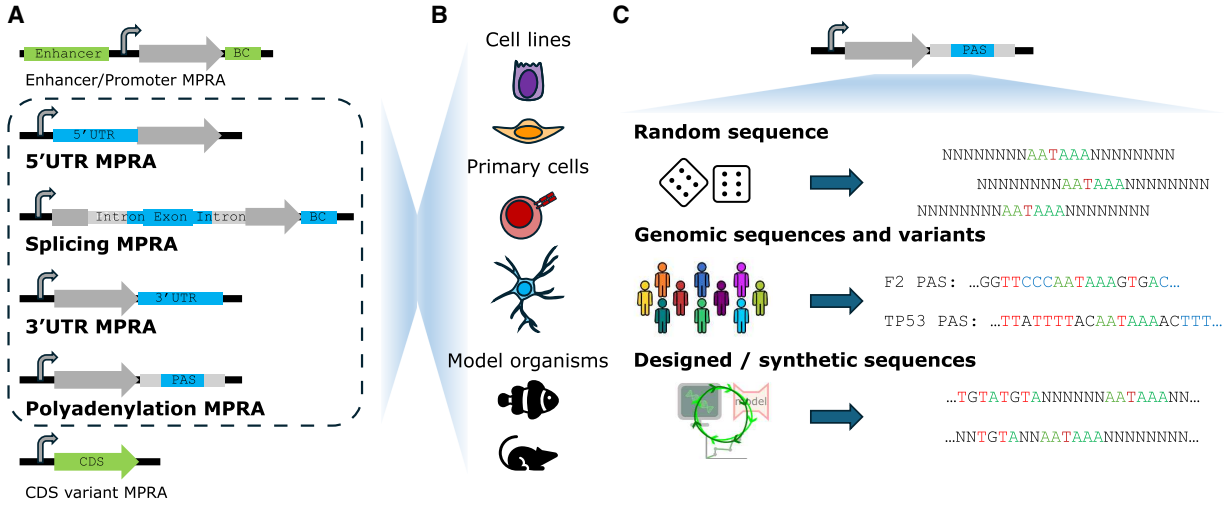

Figure 1. MPRA formats: sequence variants. (*A*) Targeting sequence variation to a specific part of a reporter gene makes it possible to comprehensively map the relationship between sequence variation and molecular phenotype. In this review, we focus on MPRAs designed to characterize the impact of variation in the 5′ and 3′ UTR sequences on stability and translation. Moreover, we discuss the MPRAs aimed at understanding the *cis*-regulatory codes controlling alternative splicing and polyadenylation. We only tangentially discuss enhancer or promoter MPRAs or assays focused on measuring variant impact on protein function and folding. (*B*) MPRAs are being performed in a wide variety of settings, but a majority of the work discussed here has been performed with human cell lines and, to a lesser extent, yeast cells. However, we also highlight work with primary cells and model organisms but cannot cover the field of bacterial MPRAs. (*C*) Sequence variation tested in MPRAs can come from a range of sources including random sequences, genomic common and variant sequences, or sequences that were designed manually or using design algorithms.

1A). We explain workflows developed in each context and discuss major experiments and results. We less comprehensively review enhancer and promoter MPRAs or those aimed at characterizing CDS variants that influence protein folding, stability, and interactions, given that many excellent reviews are available for these topics (Kinney and McCandlish 2019; Trauernicht et al. 2020; Gallego Romero and Lea 2023). Still, we highlight innovations such as single-cell MPRAs, in vivo MPRAs, or MPRAs for studying position effects introduced for enhancer and promoter analysis that provide a blueprint for similar experiments characterizing other processes. We concentrate on MPRAs exploring highly diverse sequence spaces rather than those performing saturation mutagenesis in a few gene contexts. Our focus is on MPRAs exploring human gene regulation, and we emphasize the utility of MPRA data sets for training ML models. We provide a detailed discussion of the practical applications and potential medical relevance of approaches that combine MPRAs and ML. In particular, we discuss the utility of such methods for variant stratification through variant impact measurement and prediction. We then argue that combining models and sequence design algorithms provides a powerful framework for generating synthetic sequences with the potential to improve the performance of mRNA and gene therapies.

Principles and methodology of MPRA design

This section describes key elements and underlying principles in designing an MPRA study. Here we use extensive examples to illustrate how these principles are implemented in different contexts.

Designing sequence variants

MPRAs can be broadly categorized by the type of sequence variation tested, with the two extremes being fully random sequences and genomic sequence fragments. Variation can also be generated by random or saturation mutagenesis of natural sequences, by testing specific genetic variants occurring in the human population, or by designing synthetic sequences (Fig. 1C). Many MPRAs, including some referenced in this section, do not neatly fall into one category or the other but instead test multiple types of sequences, such as a library of genome-derived fragments together with performing saturation mutagenesis for a subset of the tested sequences.

Completely random sequences are often used, providing a large-scale and unbiased interrogation of the sequence space (Ke et al. 2011; Noderer et al. 2014; Rosenberg et al. 2015; De Boer and Taipale 2024). Construction costs for random sequence libraries are low, with sequencing cost becoming the main limitation. Experiments routinely test millions of reporters (Rosenberg et al. 2015; Bogard et al. 2019; De Boer et al. 2020). Even so, as degenerate sequence length increases, it becomes impossible to screen all possible *n*-mers. The premise of randomized approaches is that *cis*-regulatory codes have a vocabulary of "words" (i.e., CREs) whose meanings and syntax can be learned if they are encountered in many different contexts.

Alternatively, MPRAs aiming to uncover *cis*-regulatory codes can screen genomic sequence fragments. For example, sequences can be selected because their accessibility profiles make them putative enhancers (White et al. 2013), because high conservation suggests a regulatory function (Oikonomou et al. 2014), or because they derive from a specific gene element (e.g., the 5′ UTR) (Zhao et al. 2014). Genomic DNA can also be randomly fragmented to create genome-covering libraries to screen for a specific regulatory function (Arnold et al. 2013; Van Arensbergen et al. 2017). Such libraries can reach the genome scale (e.g., screening all possible 5′ UTRs annotated in the genome) and are likely enriched for biologically "meaningful" sequences due to their genomic origin. However, due to natural selection, genomes—and consequently genome-derived MPRAs—are depleted for content that negatively affects survival, resulting in blind spots in the sequence to function mapping.

A third, closely related, class of MPRAs uses sequences derived from common genomic variants (Kinney et al. 2010; Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012); these assays typically compare the molecular phenotype of a sequence fragment containing the variant with that of a reference sequence. Alternatively, through random or designed mutagenesis of specific genes or CREs, MPRAs may test variants not yet observed in sequencing studies to flag putative high-impact variants or map the *cis*-regulatory information encoded in a specific gene. MPRAs focused on variants in CREs are closely related to deep mutational scanning (DMS) assays, which tend to focus on coding sequences and variants that disrupt protein folding, structure, or function but similarly quantify the impact or variants at a high level of saturation. *Cis*-regulatory MPRA and DMS workflows are sometimes grouped together as multiplexed assays of variant effect (Starita et al. 2017; Weile and Roth 2018; Kinney and McCandlish 2019).

Finally, MPRAs have been designed to screen synthetic sequences enriched for specific *cis*-regulatory motifs or motif combinations to test hypotheses about the importance of motif multiplicity, distance, or orientation (Sharon et al. 2012; Verfaillie et al. 2016; Cottrell et al. 2018; Vainberg Slutskin et al. 2018). Alternatively, synthetic sequences may be generated to achieve specific function rather than motif content (Cuperus et al. 2017; Bogard et al. 2019; Sample et al. 2019; De Almeida et al. 2024; Taskiran et al. 2024). Sequences can be designed by models or manually, but we expect generative AI to come to dominate the design of synthetic sequences for testing in MPRA formats. Synthetic sequence libraries can easily reach beyond genomic limits: Even in the simple case of manual motif embeddings, it is easy to see how variation in motif combinations, multiplicity, distance, or sequence context can result in very large libraries.

Assay formats

Many MPRAs have been described with variation targeted at different gene regions (enhancer, core promoter, 5′

UTR, exon, intron, and 3′ UTR) and designed to interrogate different regulatory processes (transcription, translation, splicing, stability, and cleavage and polyadenylation) (Fig. 1A). Although each combination of region and process requires adapting the experimental workflow, a few general assay formats have proven widely applicable. Here, we review key technological concepts, while applications to understanding gene regulation are discussed in "High-Throughput Data Analysis, Modeling, and Model Applications."

Quantifying gene expression by flow sorting and DNA sequencing The first group of assays relies on fluorescence-activated flow sorting followed by DNA sequencing (Fig. 2A; Kinney et al. 2010; Sharon et al. 2012). In these MPRAs, variation is targeted to a fluorescent reporter gene, and the reporter library is delivered to cells. Cells are sorted into bins based on fluorescence, DNA in each bin is sequenced, and each library member's activity is inferred from its distribution across the bins. Reporters are often integrated into a host cell at a single copy number because in that case, cellular fluorescence is proportional to the activity of the specific integrated construct (Noderer et al. 2014; Oikonomou et al. 2014; Zhao et al. 2014; Chong et al. 2019). Single copy integration is likely necessary when working with mammalian cells, which can take up hundreds of plasmid upon transient transfection. However, flow sorting-based MPRAs have been performed successfully with multicopy plasmids in bacteria and yeast (Kinney et al. 2010; Sharon et al. 2012), presumably because the contribution from each individual sequence can still be estimated correctly if only limited plasmid mixing occurs. A second fluorescent protein is sometimes used as a reference to correct for cell size variation and similar reporter-independent effects. In different contexts and with minor modifications, this workflow has been referred to as flow-seq (Kosuri et al. 2013), FACS-seq (Noderer et al. 2014), or sort-seq (Peterman and Levine 2016). Below, we use the term flow-seq to refer to all assays of this type.

The need for genomic integration in mammalian flow-seq assays can make flow-seq more time-consuming than workflows relying on transient DNA delivery but provides the advantage that the reporter gene is chromatinized akin to an endogenous gene. By design, this assay type connects DNA sequence to reporter protein levels and can be adapted to interrogate processes from transcription to protein stability. However, the assay provides an aggregate measurement of gene expression, requiring additional work to pinpoint the observed variation effects

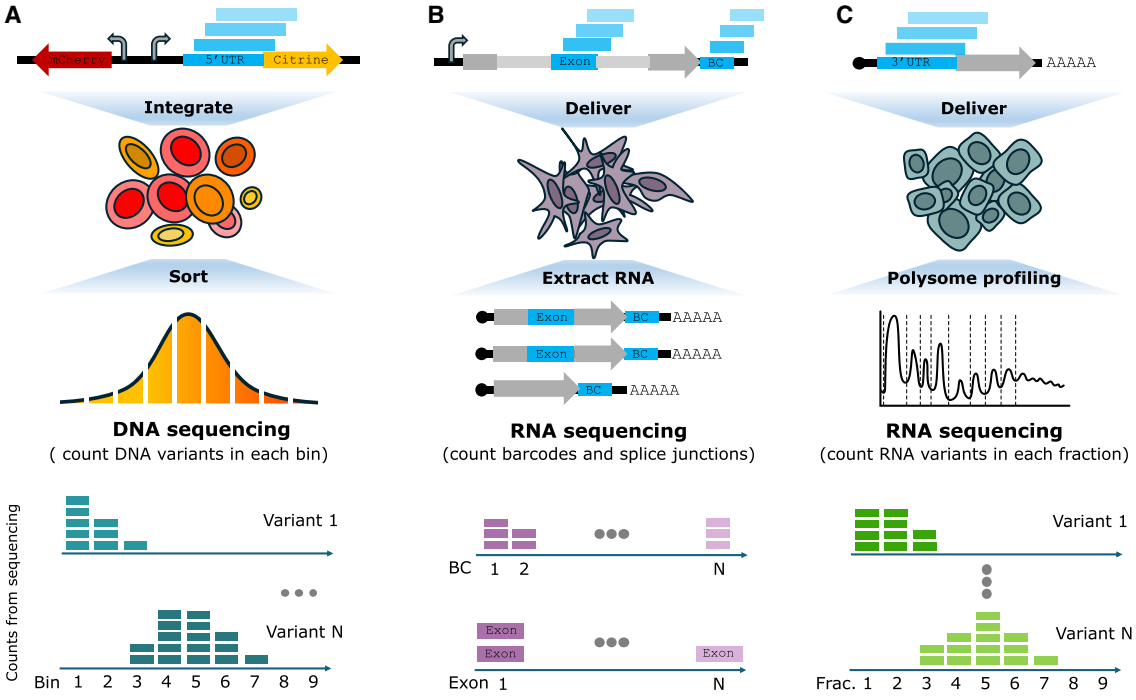

Figure 2. MPRA formats: assay variants. (*A*) Flow-seq MPRA. In the example, variation is targeted to the 5′ UTR of a reporter gene. A single copy of the DNA construct is often integrated into each cell such that fluorescence intensity is proportional to the activity of that construct. Cells are then sorted based on the normalized fluorescence (citrine/mCherry ratio), and DNA in each bin is sequenced. The activity of each variant is inferred from its distribution across the bins. (*B*) RNA-seq MPRA. The example shows a splicing MPRA where variation is targeted to an alternative exon. A 3′ UTR barcode is associated with each variable exon. After RNA extraction, sequencing across the barcode and an intron–exon junction can be used to map each transcript back to a reporter gene of origin and determine the splice isoform. (*C*) RNA-seq MPRA to quantify translation. In the example, variation is targeted to the 5′ UTR, and the library is delivered as in vitro transcribed (IVT) RNA. Centrifugation in a sucrose gradient is used to fractionate mRNA according to the number of ribosomes occupying them. Sequencing of 5′ UTRs found in each fraction is used to estimate the potential of each sequence for ribosome loading.

to a specific step. For example, several MPRAs have explored the impact of 5′ UTR variation on gene expression (Dvir et al. 2013; Noderer et al. 2014). Although many findings extracted from these data sets (for example, the strong dependence of reporter activity on out-of-frame start codons and upstream open reading frames) suggest translation modulation as a major source of the observed expression variation, 5′ UTR sequences can also modulate mRNA stability or create new transcription start sites that may affect reporter activity.

A closely related workflow uses cell growth rather than fluorescence as a readout (Kim et al. 2013; Liachko et al. 2013). In this setting, variation is targeted to an auxotrophic marker gene, and fitness is proportional to marker expression. Strongly expressed constructs result in faster growth under selection, and individual construct fitnesses can be obtained by comparing their frequency before and after selection.

Counting transcripts with RNA-seq The second broad class of MPRAs uses RNA sequencing to read out transcript or isoform abundance (Fig. 2B; Patwardhan et al. 2009, 2012; Ke et al. 2011; Melnikov et al. 2012). Transcript sequences do not necessarily contain information about all regulatory elements present at the DNA level; variation in upstream enhancers, promoters, introns, and alternative exons may not be captured. Because of this, a separate barcode sequence is often used to map transcripts to a reporter gene of origin. Barcodes can be random or designed and are associated with the regulatory sequence of interest through sequencing or DNA synthesis. Multiple barcodes are often associated with each regulatory sequence to ensure that the observed effect is not an artifact of the barcode sequence, which might contain information modulating transcript levels (Wissink et al. 2016). Of course, such redundancy also increases the library size and, therefore, synthesis and sequencing costs. The strength of each regulatory element is determined by counting the associated RNA barcodes. These counts are normalized to counts of the corresponding reporters (enhancer MPRAs), counts of an alternative isoform (alternative splicing MPRAs), or counts of the same barcode at different time points (stability MPRAs). Because all quantities of interest in RNA-seq-based MPRAs are counted directly by sequencing, these MPRA workflows do not require copy number control and are compatible with transient delivery of plasmids or with random integration using lentiviral delivery. Moreover, for stability or translation MPRAs, libraries of in vitro transcribed (IVT) RNA can be used (Sample et al. 2019), making such experiments compatible with chemically modified mRNA. Below, we discuss the specifics of MPRAs designed for quantifying splicing, polyadenylation isoform abundance, and mRNA stability.

Translation MPRAs based on polysome profiling are RNA-seq MPRAs with conceptual similarities to flow-seq (Fig. 2C; Cottrell et al. 2018; Sample et al. 2019; Niederer et al. 2022). Built on work aimed at characterizing translation of native transcript isoforms (Sterne-Weiler et al. 2013; Floor and Doudna 2016), reporter transcripts are stratified by ribosome occupancy through centrifugation in a sucrose gradient, and fractions (ribosome-free, monosome, two ribosomes, etc.) are collected and sequenced. Metrics of translational efficiency can be calculated for all sequences from counts of each transcript in each fraction. Currently, there is no common standard for which fractions to collect and how to calculate translation metrics, which can make comparison between experiments difficult. Workflows may also be sensitive to reporter length and the presence of out-of-frame (OOF) stop codons in the CDS that halt (or do not halt) translation initiated at upstream OOF start codons.

High-throughput data analysis, modeling, and model applications

One of the first challenges in learning regulatory information from MPRA data is to account for measurement noise, which can either impact each reporter activity measurement individually (e.g., variation introduced during PCR amplification, sequencing errors, etc.) or result in global differences between replicates (e.g., different sequence coverage, transfection efficiency, growth conditions, etc.). Below, we briefly reference approaches for MPRA data error correction and normalization.

A second challenge in learning from MPRA data is that their scale may allow the detection of higher-order *cis*-regulatory relationships beyond simple motif enrichment if they are present. However, doing so is not straightforward. Thus, we cover model-guided MPRA data analysis. Training ML models on MPRA data has become popular because these models can learn complex data relationships and can be used to predict activities even for sequences not yet tested experimentally. Moreover, models can be applied to stratify rare or de novo variants, guide the design of synthetic sequences, or be coupled with interpretation methods to provide insights into CRE relationships.

Error correction and model-free analysis

Proper error correction and normalization are key to learning regulatory information from MPRAs. The exact normalization and error correction details depend on the assay structure, though there are commonalities. Generally, flow-seq MPRAs count the sequences per bin and normalize each sequence's counts relative to the bin's total (Mikl et al. 2019; May et al. 2023). These per-bin values can be averaged together for a single number. RNA-seq MPRAs require comparing DNA and RNA counts or counts of different isoforms. Differential expression analysis programs such as DeSeq2 and edgeR can process RNA-seq data, with other in-depth reviews covering these pipelines (Rosati et al. 2024). Typically, these programs compare RNA expression between two conditions (e.g., no drug and drug), whereas in RNA-seq MPRAs they are used to compare, for example, DNA and RNA barcode counts. However, some of the key assumptions of these differential analysis programs, such as that most features

will not be differentially expressed, may not hold true for MPRAs. MPRA-specific (Ashuach et al. 2019; Myint et al. 2019; Gordon et al. 2020; Letiagina et al. 2021) and closely related DMS-specific (Rubin et al. 2017; Faure et al. 2020) programs for normalization and error correction programs have been developed for RNA-seq MPRAs. Still, custom approaches remain common, and a lack of standardization can complicate tasks like merging MPRA results.

After normalization and averaging as necessary, analysis for common motifs in sequences can be carried out. Quantifying effect sizes associated with specific *n*-mers provides an intuitive approach to identifying putative CREs (Ke et al. 2011; Rosenberg et al. 2015). Alternatively, sequence motifs represented as position weight matrices (PWMs) can be scanned against an MPRA library to create putative links between known *trans*-acting regulators and their impact (Kheradpour et al. 2013). Finally, postprocessing MPRA data can be added to databases of MPRA results, such as MPRAbase (Zhao et al. 2023a).

Modeling MPRA data

Early work in learning *cis*-regulatory codes from MPRA data fit equations, often inspired by biophysics, to observed trends in data (Sharon et al. 2012; Mogno et al. 2013). Classical ML models using sequence features (such as k-mer counts) as input were also common, including linear and logistic regression models (Melnikov et al. 2012; Noderer et al. 2014; Rosenberg et al. 2015; Shalem et al. 2015; Shen et al. 2016), and decision trees (Soemedi et al. 2017). At times, such relatively simple ML models can effectively capture the behavior being investigated and lend themselves to easy interpretability by examination of model weights. Regression remains a relevant modeling technique for investigating processes from stability (Leppek et al. 2022) to splicing determinants (Chiang et al. 2022). Likewise, ensemble models of decision trees remain popular, such as using gradient-boosted regressors for predicting splicing (Mikl et al. 2019) and mRNA stability and localization (Mikl et al. 2022) or using random forests for predicting 5′ UTR effects on protein production (Cao et al. 2021).

With their many parameters, deep learning models are adept at modeling nonlinearities. The development of neural network models using DNA sequences as inputs (Alipanahi et al. 2015; Kleftogiannis et al. 2015; Zhou and Troyanskaya 2015) has paved the way for these approaches to be applied to MPRA data (Cuperus et al. 2017; Paggi et al. 2017; Bogard et al. 2019; Cheng et al. 2019; Movva et al. 2019; Sample et al. 2019; Vainberg Slutskin et al. 2019). Deep learning models for genomics have been reviewed in detail elsewhere (Eraslan et al. 2019; Zou et al. 2019), but we briefly cover a few key concepts. A network architecture often used with MPRA data is the convolutional neural network (CNN) (Fig. 3A, left). CNNs include convolutional layers, which consist of pattern-detecting filters that are scanned across inputs to evaluate how well each position matches the filter. The final layers in a CNN compress prior layer information into a set number of outputs. Recent MPRA modeling work has exploited architectural changes to enhance model performance and interpretability. For example, deeper networks and layer-skipping connections have improved alternative polyadenylation (APA) and promoter activity predictions (Linder et al. 2022a; Penzar et al. 2023). By structuring networks around our knowledge of processes like splicing, they can be forced to be interpretable, and testable hypotheses

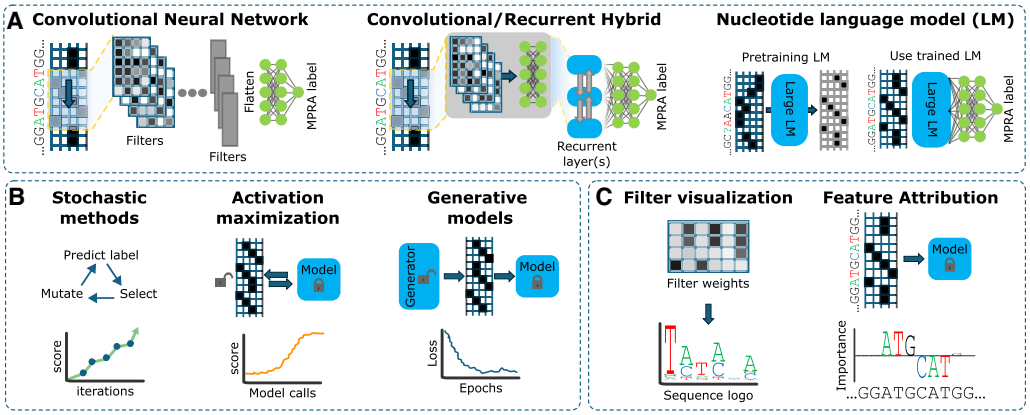

Figure 3. Models and sequence design. (*A*) CNNs have become widely used to model MPRA data. The convolutional filters in the first layer are scanned along the one hot encoded sequence. Additional convolutional layers help capture nonlinear interactions between the filters in the first layer. CNN/recurrent neural network (RNN) hybrid networks have been developed to handle longer or variable input sequences. They typically have CNN-like layers followed by recurrent layers. Nucleotide language models (LMs) are very large unsupervised models trained to reconstruct input sequences. Once trained, layer(s) can be removed and connected to a predictor network instead, before supervised training on labeled data. (*B*) Sequence design algorithms. For these, it is assumed that the design goal is to maximize some score. The lock symbol corresponds to models with frozen weights or models/sequences that are free to update their values and weights. (*C*) Approaches to neural network interpretation. In filter visualization, the weights of a convolutional filter can be represented like a PWM, which can be visualized as a sequence logo. In feature attribution, an attribution method is used to determine numeric values for the importance of each feature in an input sequence to a pretrained model.

about RNA features can be extracted from models (Liao et al. 2023; Gupta et al. 2024).

Initially, CNNs trained on MPRA data were designed to operate on inputs matching the size of the varied regions in an assay. However, given the premise that what is learned about regulatory processes using MPRAs extends beyond said assays, applying MPRA-trained models to longer sequences is attractive. Multiple operations and network structures can be used for this goal. For example, global pooling operations can summarize CNN layers into fixed-length vectors regardless of input lengths. Such operations have been used to train a model on translation MPRA data to predict the ribosome loading of native transcripts (Karollus et al. 2021). Non-CNN architectures like recurrent neural networks (RNNs), which accept long inputs, can also be used. In RNNs, information moves sideways through layers and forward from input to output, allowing recurrent layer nodes to have internal memory states that allow the preceding and following nucleotides in a sequence to influence the current state. Genomic models often also combine CNN and RNN elements (Fig. 3A, middle; Quang and Xie 2016; Angermueller et al. 2017). When training multilength predictors, data are often combined from multiple sources, including MPRAs (Agarwal and Kelley 2022; Li et al. 2022).

A final class of models becoming relevant to MPRAs is large language models (LLMs) (Fig. 3A, right). LLMs have yielded promising results in protein sequence-to-function prediction tasks (Rives et al. 2021; Lin et al. 2023), where they are trained on databases of protein sequences to fill in masked positions with the most likely amino acid. These models can then be used to generate sequence embeddings for a downstream model, with the assumption that the LLM will already contain protein sequence distribution knowledge to build off of (Biswas et al. 2021). Similar foundational LLMs have begun to be trained on genomic data (Ji et al. 2021; Consens et al. 2023; Karollus et al. 2024), with some geared toward specific regions such as 5′ UTRs (Chu et al. 2024) or splicing sites (Chen et al. 2024). However, recent work found that pretraining schemes and training data sets can heavily affect genomic LLM performance (Tang and Koo 2024; Vilov and Heinig 2024).

Sequence design using MPRA-trained models

The most direct approach to model-guided sequence design is to perform an in silico evolution experiment (Fig. 3B, left) where a sequence is iteratively mutated, with mutations resulting in predicted activity values closer to a target value being kept. This approach has been used to engineer 5′ UTRs for yeast (Cuperus et al. 2017) and human genes (Sample et al. 2019; Cao et al. 2021), and enhancers for *Drosophila* cell lines (De Almeida et al. 2022). However, stochastic search can be computationally expensive because many mutations do not result in higher activity but still need to be scored before rejection.

A more efficient approach for differentiable predictors like neural networks is optimizing the input pattern by gradient ascent, an approach also termed "activation maximization" (Fig. 3B, middle). Gradient ascent is often used in machine learning to numerically find the local maximum of a continuous and differentiable function: Rather than randomly exploring the coordinate space, each step is taken in the direction of the function's largest increase. Because nucleic acid sequences are discrete, they must be approximated by a continuous representation to enable gradient ascent (Lanchantin et al. 2016; Killoran et al. 2017). This approximation often resembles a PWM where each nucleotide occurs with some probability per position. Initially introduced for sequence visualization (Lanchantin et al. 2016), a variety of improvements have been made to this approach to enable fast design of DNA and protein sequences (Killoran et al. 2017; Bogard et al. 2019; Linder and Seelig 2021; Norn et al. 2021). Gradient design algorithms have been used with MPRA-trained models to generate various regulatory sequences (Bogard et al. 2019; Gosai et al. 2023; Castillo-Hair et al. 2024). Still, activation maximization, like stochastic search, is not guaranteed to find globally optimal sequences.

Generative approaches like generative adversarial neural networks (GANs) (Goodfellow et al. 2014) and variational autoencoders (VAEs) (Kingma and Welling 2013) have been adapted to sequence design (Killoran et al. 2017; Brookes and Listgarten 2018; Costello and Martin 2019; Gupta and Zou 2019; Linder et al. 2020; Repecka et al. 2021; Shin et al. 2021; Zrimec et al. 2022; Uehara et al. 2024). In these approaches, a pretrained predictor is used to train a separate model that learns to generate sequences that maximize a target, often with additional penalties to prevent generated sequences from becoming too diverged from predictor training sets (Fig. 3B, right). Although training generators can be computationally costly, sequences can be efficiently generated once training is complete. Additionally, genomic autoregressive LLMs can generate sequences for tasks like promoter and enhancer design (Lal et al. 2024). In autoregressive LLM generation, a sequence is built by choosing nucleotides to lengthen it based on the existing sequence and the information about sequence space that the LLM has learned. Diffusion models have also recently gained traction for the design of synthetic regulatory elements. Diffusion models are generative models where noise is added to the training data and the model learns to "denoise" the data and recover the underlying information (Avdeyev et al. 2023; Penzar et al. 2023; DaSilva et al. 2024; Sarkar et al. 2024; Stark et al. 2024). For generation, the model is fed randomly generated inputs, which it denoises to create sequences that should be similar to the training data and thus are plausibly regulatory elements.

Finally, we note that design quality, regardless of the algorithm used, is constrained by the quality of the predictor. Common approaches to attempt to improve a model for specific *cis*-regulatory processes involve collecting more training data. For example, sequences that align with a specific design task can be generated, assayed with an MPRA, and used to update a model. An example of this iterative approach used base models trained on previous enhancer MPRA and chromatin accessibility data to generate initial designs for cell type-specific enhancers,

which were measured with an MPRA. These initial designs were used to update the models, and a second round of dramatically improved cell type-specific enhancer designs was created (Yin et al. 2024). Active learning offers an alternative guide for collecting data to improve general model performance. In active learning, a base predictor is iteratively improved by measuring data points with uncertain model predictions. For MPRA modeling, this involves a cycle of generating sequences to assay, measuring them with an MPRA, and using the new data to update the model. Recently, active learning was used to train a classifier of enhancer/silencer activity for sequences containing binding motifs for the retina-specific TF cone–rod homeobox. Starting from a near-random classifier trained on assayed genomic sequences and a set of likely disruptive mutants, by round four of active learning, top classifier performance nearly doubled (Friedman et al. 2023). Additionally, recent work exploring how data set size, diversity, and model complexity relate when training *cis*-regulatory predictors found that the amount of data needed for training CNNs may be similar to that needed to train simple models in some cases (Nikolados et al. 2022).

Interpreting MPRA trained models to learn cis*-regulatory rules*

The nonlinearities that make neural networks so computationally powerful make it challenging to understand how they arrive at their predictions. Methods for interpreting genomic regulatory predictors have been reviewed elsewhere (Novakovsky et al. 2023), but we summarize relevant approaches here. For CNNs, it is possible to visualize what filters have learned using their weights, resulting in sequence motifs similar to position weight matrices, which sometimes match known *cis*-regulatory motifs or parts of known motifs (Fig. 3C, left; Alipanahi et al. 2015). Visualization has been used to create filter PWMs for the first-layer filters in many MPRA-trained CNNs (Cuperus et al. 2017; Vainberg Slutskin et al. 2019; Park et al. 2022; Klie et al. 2023; Reimão-Pinto et al. 2023). It is possible to generalize this approach to visualize filters in deeper CNN layers, some of which have been mapped to combinations of motifs (Bogard et al. 2019). By manipulating CNN training (Koo et al. 2019), structure (Koo and Eddy 2019), and activation functions (Koo and Ploenzke 2021), the filters of shallow CNN layers can be encouraged to be more motif-like for visualization.

Understanding what models have learned beyond motifs and motif combinations extracted from CNN filters is critical for deciphering more complex regulatory grammars. Neural network interpretation techniques that assign values to features representing how much they contribute to a network prediction for each sequence were developed to address this limitation (Fig. 3C, right; Shrikumar et al. 2019; Carter et al. 2020; Lundberg et al. 2020; Avsec et al. 2021; Linder et al. 2022b). Interpretation methods can be useful in comparing wild-type and variant *cis*-regulatory sequence predictions, potentially finding features beyond the mutated position(s) contributing to the prediction (Minnoye et al. 2020).

Although examining solitary sequence attributions is useful, attribution methods focusing on relationships between important features across a data set can help discover new regulatory rules (Greenside et al. 2018; Wei et al. 2023; Seitz et al. 2024). One such tool is TF-MoDISco, which searches for common motifs in a data set using model importance scores (Shrikumar et al. 2018). These attribution-based motifs can then be examined for distance-dependent or cooperative effects (Avsec et al. 2021; Agarwal and Kelley 2022). For example, a CNN trained to predict fly enhancer activity used TF-MoDISco to find motifs associated with development or housekeeping enhancers and then examined the effects of flanking sequence and distance between motifs on predicted activity (De Almeida et al. 2022). These intermotif patterns were validated by screening sequences with an MPRA swapping motif flanking nucleotides and differing distances between motifs.

Learning *cis*-regulatory codes governing RNA processing, stability, and translation

This section focuses on MPRAs for learning *cis*-regulatory codes governing splicing, polyadenylation, stability, or translation from variations targeted to 5′ UTRs, exons, introns, and 3′ UTRs. MPRAs aimed at identifying disease-relevant variants will be discussed in a later section.

Learning how the 5′ UTR sequence modulates translation

Regulatory elements in the 5′ UTR are the main determinants of mRNA translation initiation. Translation begins when the 43S preinitiation complex (PIC) is recruited to the 5′ cap. The PIC then scans the 5′ UTR until it encounters a start codon, commonly AUG, where ribosome assembly is finished (Sonenberg and Hinnebusch 2009). Sequence elements in the 5′ UTR alter the scanning rate and start codon recognition. Upstream start codons are present in 50% of human 5′ UTRs (McGillivray et al. 2018) and compete with the start codon of the main open reading frame for the PIC. Secondary structure also affects translation initiation, with strong stem–loops negatively impacting translation (Leppek et al. 2018). The role of 5′ UTRs in modulating mRNA stability is not fully understood, and CDS and 3′ UTR elements may be more important in determining stability (Agarwal and Kelley 2022). Still, mRNA translation and stability are tightly coupled, and the 5′ UTR sequence is likely to impact mRNA stability at least indirectly by modulating translation (Wu and Bazzini 2023).

MPRAs have contributed substantially to furthering our knowledge of these processes, especially by enabling us to quantitatively characterize the impact of specific CREs. Two foundational reports studied how the sequence context around a start codon modulates expression using flow-seq MPRAs in yeast (Dvir et al. 2013) and human cell lines (Noderer et al. 2014). They quantified requirements for efficient reporter expression,

including the importance of a purine at the −3 position and the negative impact that out-of-frame (OOF) start codons have on reporter fluorescence. These observations agree with the notion that translation regulation is the primary, though not only (Dvir et al. 2013), mechanism responsible for the observed fluorescence variation. Similarly, MPRAs were made to explore the impact of upstream open reading frames (uORFs) on reporter expression (Lin et al. 2019; Jia et al. 2020; May et al. 2023). By randomizing a 10 bp stretch upstream and overlapping the start codon of a short uORF, an IVT mRNA MPRA found that RNA stability correlated negatively with uORF translation and positively with CDS translation (Jia et al. 2020). An MPRA of uORF-containing native yeast 5′ UTRs found that uORFs with an AUG start are more repressive than those with non-AUG starts (May et al. 2023). Additionally, MPRAs have enabled the discovery of CREs modulating translation. An in vitro translation MPRA of native yeast 5′ UTRs identified C-rich motifs as translation inhibitors while validating U-rich sequences as translation enhancers (Niederer et al. 2022).

Further advancements in understanding 5′ UTRs' roles in translation regulation have come from modeling MPRA data. Dvir et al. (2013) trained a linear regression model on MPRA data that explained two-thirds of the observed protein level variation using relatively few features. However, the small scale and narrow sequence variation window of the data set limited the model's applications (Dvir et al. 2013). To learn models of the *cis*-regulatory code beyond the vicinity of the start codon, Cuperus et al. (2017) randomized 50 bp upstream of the start of a yeast auxotrophic marker and quantified the fitness of 500,000 5′ UTRs using growth selection followed by sequencing. Sample et al. (2019) performed a randomized 5′ UTR MPRA of similar scale in human cell lines using IVT RNA and a polysome profiling RNA-seq assay. CNNs trained on these data sets predicted the translation impacts of native (yeast or human) 5′ UTR fragments, suggesting that the underlying *cis*-regulatory grammar is learnable from random sequences alone. Filter visualization supported this, as some filter PWMs matched known translation-affecting motifs. Using the same data, Karollus et al. (2021) introduced a reading frame-sensitive CNN architecture with global pooling capable of predicting the translation effects of human 5′ UTRs of arbitrary lengths.

As the availability of MPRA 5′ UTR translation efficiency data sets increased, models began to be trained on combined data sets. For example, Zheng et al. (2023) trained a model on human and yeast MPRA data, which yielded improved predictions on the yeast 5′ UTR data set from Cuperus et al. (2017). LLMs are trained on large, multispecies data sets and have been used for inputs to models trained on MPRA 5′ UTR data that performed better than simpler models (Chen et al. 2022; Chu et al. 2024).

Multiple MPRAs investigating cell type-specific 5′ UTR translation effects have found few differences between different cell lines or activated T cells (Ferreira et al. 2013; Noderer et al. 2014; Lim et al. 2021; Castillo-Hair et al. 2024). Polysome profiling and sequencing experiments used to characterize the differentiation of hESCs into neurons support the robustness of 5′ UTR-mediated translation regulation to different cellular environments (Blair et al. 2017). Still, these MPRAs were done in proliferating, efficiently translating cells. Different translation regulation patterns might be observed in nonproliferating cells or in cell states characterized by different ribosome levels or composition. Two recent studies reported on using IVT mRNA MPRAs to understand translation regulation in developing zebrafish embryos (Strayer et al. 2023; Reimão-Pinto et al. 2023). Both studies observed differential regulation of translation between subsets of 5′ UTRs derived from maternal and zygotic transcripts, suggesting dynamic regulation of translation during development.

Deciphering the cis*-regulatory code of alternative splicing*

Alternative splicing (AS) is a major source of proteome diversity and is regulated by *cis*-regulatory sequences present in pre-mRNA and the *trans*-acting RNA binding proteins (RBPs) that recognize them (Wright et al. 2022; Marasco and Kornblihtt 2023; Rogalska et al. 2023). The core splicing signals, the 5′ splice donor (SD), 3′ splice acceptor (SA), branch point, and polypyrimidine tract are required to recognize intron–exon boundaries and intron removal. The next level of regulation is formed by splice regulatory elements (SREs) in exons or introns recognized by cognate RBPs.

Alternative splicing is well suited for RNA-seq MPRAs because targeted sequencing can easily determine the identity and abundance of isoforms of interest. Although details vary, AS MPRAs typically target variation to intronic and exonic regions expected to modulate usage of nearby splice sites. Sequencing across exon–exon or exon–intron boundaries reveals the splice isoform of each transcript. A 3′ UTR barcode can link transcripts to reporter genes even when the variable exon or intron is spliced out (Rosenberg et al. 2015). Alternatively, spliced-in read counts can be normalized to input DNA counts to estimate splicing efficiency (Ke et al. 2011).

The use of random sequences to identify SREs at scale has a long tradition in the splicing field (Wang et al. 2004; Yu et al. 2008; Culler et al. 2010). Ke et al. (2011) quantified the impact of all hexamers in five different positions in two exon-skipping reporters. A similar approach was taken by Wong et al. (2018), who tested all possible 5′ SD sequences in three different exon-skipping contexts. Findlay et al. (2014) used CRISPR/Cas9-mediated genome editing and homology-directed repair to substitute a hexamer in exon 18 of BRCA1 in a haploid cell line with all possible hexamers. They observed strong effects on the RNA/DNA ratio due to nonsense-mediated decay and the insertion of exonic SREs (Findlay et al. 2014). These experiments both succeeded at characterizing the impact of all *n*-mers in different splicing contexts and highlighted context dependencies, including the impact of neighboring sequences and the importance of position relative to the SD and SA. Overall, they suggest that it is not always

possible to capture *n*-mer impacts on AS by a single, universal score.

To generalize beyond single motifs in fixed contexts, Rosenberg et al. (2015) created reporter libraries with more extensive sequence variation (25 nt) targeted to an alternative exon between two competing SDs or two competing SAs. Over 2 million reporters were assayed in human cell lines, and a linear regression model was trained on these data. Although the model was trained on 5′ and 3′ AS, it generalized to predict exonic variant impact on cassette exon inclusion, supporting the idea of a universal exon definition code shared between different types of AS. Rather than generalizing predictions from other contexts, Regev and colleagues (Liao et al. 2023) developed a large-scale random MPRA to directly characterize the impact of exonic sequence variation on cassette exon inclusion. A CNN with an architecture constrained to promote interpretability performed well using only 13 filters, most of which could be mapped to known splicing regulatory motifs.

To generalize splicing beyond exonic variants, Mikl et al. (2019) tested panels of endogenous and synthetic splice regulatory sequences (e.g., SDs and SAs) in the contexts of intron retention, exon skipping, 5′ AS, and 3′ AS reporters and trained a gradient-boosting regression model that showed generalization ability to other data sets. Other work focused on combining data sets to improve splicing predictions. Gagneur and colleagues (Cheng et al. 2019) trained a modular model architecture on MPRAs and other data sources . A later version of this modular model was fine-tuned on human RNA-seq data to capture tissue-specific AS (Cheng et al. 2021). An explicitly interpretable splice prediction model trained on transcriptomic data used data from a random MPRA for validation (Gupta et al. 2024).

Similarly, work has focused on using MPRAs to study intronic variation in yeast. Schirman et al. (2021) created synthetic intron libraries through combinatorial assembly of naturally occurring splice site variants to study constitutive splicing in yeast. They found that splicing efficiency is positively correlated with RNA abundance (Schirman et al. 2021). Complementary work used a large library of random intron sequences to characterize yeast splicing determinants, finding that over two-thirds of intron sequence decreased splicing efficiency, with high G/C content and secondary structure having a particularly negative effect (Perchlik et al. 2024).

Decoding the polyadenylation code

The 3′ ends of most eukaryotic mRNAs are formed by cleavage and polyadenylation (CPA), an essential step in mRNA maturation that directly impacts transcription termination, transcript export, translation, and stability. CPA is controlled through the interplay between CREs at the poly(A) site (PAS) and a large number of *trans*-acting factors, including the core CPA machinery and regulatory factors (Chan et al. 2011; Tian and Manley 2017; Mitschka and Mayr 2022). Mammalian mRNA PASs are typically defined by a conserved AAUAAA hexamer, a variable U/GU-rich downstream element, and other auxiliary sequences. These CREs are recognized by the core CPA factors to assemble the mRNA 3′ processing complex. CPA occurs ∼20 nt downstream from a core hexamer, but cut positions further downstream are observed in many transcripts (Gruber et al. 2016; Gruber and Zavolan 2019). Upward of 50% of human genes produce multiple mRNA isoforms by using alternative PASs (Derti et al. 2012; Hoque et al. 2013; Lianoglou et al. 2013). Different APA isoforms from the same gene can encode different proteins or be differentially regulated (Tian and Manley 2017; Mitschka and Mayr 2022).

In contrast to splicing, for which core sequences are often physically separated by long distances, the essential sequence motifs for pre-mRNA 3′ processing are located together in a short region, making APA highly amenable to study with MPRAs. Two complementary RNA-seq MPRAs were used to interrogate the *cis*-regulatory poly(A) code. The first approach tested >3 million PASs with random regions in an APA MPRA (Bogard et al. 2019). The strength of each PAS was quantified relative to a strong competing PAS. A second approach tested a library of 12,000 human and viral PASs and their variants (Vainberg Slutskin et al. 2019). Each reporter construct contained only a single PAS, and PAS strength was inferred from RNA abundance, given that only transcripts that undergo CPA are stabilized and exported to the cytoplasm.

APARENT and PolyApredictor, two CNNs trained on these data sets, accurately predicted PAS usage and cleavage position in the MPRA data and generalized to predict the strengths of independently measured human polyadenylation events (Bogard et al. 2019; Vainberg Slutskin et al. 2019). APARENT2, a deep residual CNN subsequently trained on the random APA MPRA data set, improved generalization to native transcripts and also accurately predicted the relative strength of the PASs tested in the nonalternative MPRA, suggesting that the same *cis*-regulatory code governs APA and nonalternative polyadenylation (Linder et al. 2022a). Moreover, although trained on MPRA data, APARENT2 predictions agreed well with in vitro pre-mRNA 3′ processing measurements (Liu et al. 2023). PolyaStrength, a model trained on human transcriptomic data that performed well at predicting APA MPRA data, supports the notion that *cis*-regulatory rules governing APA are shared between different contexts (Stroup and Ji 2023).

In addition to the core sequence motifs, MPRAs have been applied to investigate the contribution of auxiliary regulatory sequences to CPA. To understand the determinants of distal CPA events, Wu and Bartel (2017) created an MPRA with random sequences downstream from the core hexamer and assayed the cleavage position. Distal cuts were found to be enabled by secondary structure formation, which likely shortens the physical distance between the core hexamer and the cleavage site (Wu and Bartel 2017). These structures may also directly stimulate pre-mRNA 3′ processing efficiency, as suggested by an early in vitro mutagenesis and selection study (Graveley et al. 1996).

3′ UTR MPRAs for learning the cis*-regulatory codes governing mRNA stability and translation*

Sequence elements in the 3′ UTR can recruit RBPs and miRNAs, influencing ribosome recruitment, elongation, stability, and subcellular localization in addition to CPA (Mayr 2017). For example, Pumilio binding sites (Van Etten et al. 2012) and AU-rich elements (Barreau 2005) are both major determinants of mRNA stability (Agarwal and Kelley 2022). Regulatory elements that allow for cell type-specific translation are also found in 3′ UTRs, whereas 5′ UTRs mostly control translation at a global level (Floor and Doudna 2016).

Because of the important role that 3′ UTR elements have in regulating mRNA stability, many MPRAs have focused on quantifying the relationship between 3′ UTR sequence and mRNA stability or abundance. Stability MPRAs track the degradation of an mRNA population after a transient pulse of mRNA delivery or inhibition of transcription (Zhao et al. 2014). mRNA half-lives are commonly calculated by fitting the data to an exponential decay model, though more complex models are sometimes necessary to capture the observed decay kinetics (Rabani et al. 2017; Castillo-Hair et al. 2024).

Alternatively, mRNA stability can sometimes be inferred from steady-state measurements, assuming that the mRNA production rate is constant. Steady-state measurements are convenient because RNA (and reference DNA) can be collected just once. The assumption of a constant production rate is justifiable if all reporter constructs use an identical, strong promoter. In at least one case where such a comparison was made, mRNA steady-state abundance and half-lives were well correlated (R = 0.62), though outliers were observed (Zhao et al. 2014). Still, it is important to note that we cannot generally equate abundance with stability. The STARR-seq workflow demonstrates that sequences embedded in the 3′ UTR of a reporter construct can act as enhancers of a minimal promoter and, in that case, the primary mechanism for varying RNA abundance is through regulation of transcription (Arnold et al. 2013).

Early 3′ UTR MPRAs focused on characterizing libraries of human 3′ UTR fragments and on identifying putative CREs from (steady-state) flow-seq measurements (Oikonomou et al. 2014) or from a combination of flow sorting with an mRNA decay time course (Zhao et al. 2014). Subsequently, such workflows were adapted to quantify the impact of random 8-mer sequences (Wissink et al. 2016), AU-rich elements, and known RBP binding sequences (Siegel et al. 2022). These experiments found CREs modulating gene expression: Some mapped to known RBP and microRNA binding sites, whereas others were novel.

Measurements of mRNA stability or abundance generally correlate well with protein reporter levels measured in flow-seq MPRAs, suggesting that CREs in the 3′ UTR primarily modulate mRNA stability rather than translation (Zhao et al. 2014; Wissink et al. 2016). Still, direct translation measurements are necessary to explain the residual between measurements of mRNA stability and reporter protein expression. Recent MPRAs of human 3′ UTR fragments (Schuster et al. 2023), synthetic 3′ UTRs containing miRNA and RBP binding sites (Cottrell et al. 2018), and fragments from 143 viral genomes (Seo et al. 2023) used polysome profiling and fraction sequencing to quantify translation. Overall, these experiments found less pronounced variation in translation efficiency than in mRNA stability but still identified many CREs modulating translation. Intriguingly, Seo et al. (2023) identified CREs that enhanced both processes compared with a strong control sequence, which could have mRNA and gene therapy applications.

MPRAs performed across developmental stages have revealed global and sequence-specific stability differences. Two MPRAs investigated mRNA stability during the maternal-to-zygotic transition in developing zebrafish (Rabani et al. 2017; Yartseva et al. 2017). IVT mRNA containing native 3′ UTR fragments was injected at the 1 cell stage, and RNA was collected at later time points. These experiments revealed two distinct mRNA populations: "early" and "late" onset. Early-onset mRNA is sensitive to degradation by maternally deposited factors and decays exponentially upon IVT mRNA injection. Conversely, late-onset mRNA is stable until the zygotic degradation machinery, including miR-430 and Pumilio, are expressed (Rabani et al. 2017).

Similarly, 3′ UTR MPRAs have looked for stability regulation differences between cell lines. In MPRAs with synthetic 3′ UTRs containing high-affinity miRNA or RBP targets, varying miRNA and RBP expression levels accurately explained the observed reporter activities in cell lines (Cottrell et al. 2018; Vainberg Slutskin et al. 2018). Still, in other cell line experiments focusing on human 3′ UTR sequences, reporter mRNA expression is often highly correlated between contexts (Zhao et al. 2014; Griesemer et al. 2021). The lack of extensive cell type-specific mRNA stability regulation in native sequence libraries instead of synthetic, CRE-enriched sequences may not be surprising: Only a small subset of miRNAs or RBPs is expressed in a given cell type, and only a small subset of CREs is expected to respond to them (Alles et al. 2019; McGeary et al. 2019).

As in other contexts, 3′ UTR MPRAs are used to train and validate computational models. A linear regression model using n-mer features was trained to predict stability from sequence on zebrafish stability MPRA data and could capture both early- and late-onset decay (Rabani et al. 2017). No similar model has been trained on MPRA data from human cells, but MPRA data have been used to validate a model trained on stability measurements for native transcripts (Agarwal and Kelley 2022). The model achieved Spearman correlations of r = 0.63, r = 0.49, and r = 0.26–0.50 on MPRA data from Litterman et al. (2019), Griesemer et al. (2021), and Siegel et al. (2022). These values are worse than the model's performance on native stability test data, possibly due to differences in how stability is measured in the MPRAs.

Additionally, MPRAs have investigated the impact of 3′ UTR sequences on gene expression in yeast, with one focusing on native 3′ UTRs and their variants (Shalem et al.

2015) and the other using random sequences (Savinov et al. 2021). Both studies identified the efficiency element, similar to the human core polyadenylation hexamer, as a primary *cis*-regulatory determinant of protein expression even though neither set out specifically to study CPA (Shalem et al. 2015; Savinov et al. 2021). Recently, in modeling the random 3′ UTR MPRA data, a linear regression model trained on DNA LLM representations of sequences outperformed models trained on k-mer representations of the same sequences (Karollus et al. 2024). This success highlights the potential of LLM embeddings to help explain MPRA data and the utility of said data for validating and testing advanced *cis*-regulatory machine learning models.

Applications and medical relevance

Variant impact measurements and prediction

A practical application for MPRAs is stratifying human genetic variants. Various tools, including genome-wide association studies (GWASs) (Uffelmann et al. 2021), quantitative trait locus analysis, and conservation-based pathogenicity predictors (e.g., PolyPhen-2 [Adzhubei et al. 2013] and CADD [Kircher et al. 2014]) are available to identify putative pathogenic variants. However, these tools either rely on statistical associations and struggle with rare variants (Bomba et al. 2017), fail to separate causal from bystander variants, or cannot propose a mechanism of action.

MPRAs directly measure a variant's molecular phenotype and thus can suggest a mechanism by which it can change protein expression. In order to be transferable from the reporter setting, variant effects are quantified relative to a reference. Variant-focused MPRAs have been performed for translation, stability, polyadenylation, and splicing. Studies focused on splicing are the most common because splice-disrupting variants can significantly affect protein sequence and could explain how synonymous pathogenic variants operate.

Saturation mutagenesis of disease-relevant exons and UTRs Several AS MPRAs have characterized variant impacts on splicing in the context of disease-relevant exons. These include exon 6 of FAS/CD95 (Julien et al. 2016), WT1 exon 5 (Ke et al. 2018), RON exon 11 (Braun et al. 2018), CD19 exons 1–3 (Cortés-López et al. 2022), and exon 2 of POU1F1 (Gergics et al. 2021). Furthermore, saturation mutagenesis screens have been performed for ∼20 PASs in human disease-associated genes, including ACMG genes BRCA1, BRCA2, PTEN, and TPMT (Bogard et al. 2019), and for the CXCL2 3′ UTR (not including the PAS), which destabilizes its mRNA (Zhao et al. 2014). Such mutagenesis data sets provide clinical practitioners with lookup tables of variant impacts and are useful benchmarks for model performance, as shown in the case of splicing predictors (Smith and Kitzman 2023). However, such focused assays are ill suited for training models expected to generalize beyond their target exon or gene. We also note that this review focuses on a small subset of mutagenesis MPRAs: Many studies have been performed characterizing the impact of CDS variants on protein folding, function, or interactions (Starita et al. 2017; Weile and Roth 2018; Kinney and McCandlish 2019).

High-throughput testing of variants from ClinVar, ExAC, HGMD, and patient cohorts Complementary work has taken a variant- or disease-centric approach, testing variants from public databases such as ClinVar (Landrum et al. 2018), HGMD (Stenson et al. 2003), ExAC (Lek et al. 2016), and gnomAD (Karczewski et al. 2020) or from patient cohorts. Soemedi et al. (2017) tested almost 5000 disease-associated exonic variants in a splicing MPRA in cell-free and cell line settings. Adamson et al. (2018) screened 2059 genetic variants from ExAC occurring in 110 alternative exons. Chong et al. (2019) further scaled up variant characterization by testing the impact of almost 28,000 ExAC variants in 2198 human exons using reporters integrated at a defined locus. Chiang et al. (2022) performed an MPRA focused on intronic variants selected from ClinVar, HGMD, and similar sources occurring near branch points or in intronic regions upstream of 5′ splice sites.

Different AS MPRAs found frequencies of splice-disrupting variants ranging from 70% (Ke et al. 2011) to 3.8% of variants (Chong et al. 2019). A priori, this is unsurprising because each experiment uses a different reporter system and tests different variant classes and exons. Additionally, variants are expected to impact splicing more if the exon is commonly alternatively spliced than if the common exon is constitutively spliced (Baeza-Centurion et al. 2019; Glidden et al. 2021). We also note that MPRA isoform ratios may not represent that of the native context. Still, variant-induced change in isoform abundance in the native context can be estimated from MPRA measurements and knowledge of the isoform ratio for the reference sequence in the genomic context (Baeza-Centurion et al. 2020).

Variant MPRAs have not been limited to splicing. A screen of 3577 5′ UTR variants from ClinVar identified several that substantially impact ribosome loading, many of which inserted or deleted uORFs (Sample et al. 2019). Hsieh and coworkers (Lim et al. 2021; Schuster et al. 2023) used MPRAs to quantify the impact of thousands of 5′ and 3′ UTR variants from prostate cancer patient genomes and found many that substantially altered either translation or stability. Interestingly, analysis of the 3′ UTR data revealed that mutations in highly conserved regions were more likely to impact mRNA stability than translation and that stability-modulating variants had a larger effect on patient outcomes. An MPRA was used to characterize the impact of variants that disrupt miRNA seed sites (Ipe et al. 2018). Two recent studies used MPRAs to measure the impact of tens of thousands of 3′ UTR variants on mRNA abundance (Griesemer et al. 2021; Fu et al. 2024). Variants were selected because of their potential disease relevance: They are in strong linkage with GWAS tag single-nucleotide polymorphisms, occur in regions under positive selection, are rare

in the human population (<0.1% allele frequency in gnomAD) (Karczewski et al. 2020), or are somatic mutations in cancer-related genes from the COSMIC database (Tate et al. 2019). Complementary work characterized the impact of ClinVar, HGMD, or GWAS variants that occur in 3′ UTRs near PASs and thus potentially disrupt CPA (Bogard et al. 2019; Linder et al. 2022a). Overall, these studies identified a notable number of variants that significantly impact RNA levels or processing. For example, 38% of rare gnomAD variants were reported to be functional in at least one of two cell lines used for one of the MPRAs (Fu et al. 2024).

Finally, MPRAs and models have been used to stratify variants from the Simons Simplex Collection of genome sequencing data from autism spectrum disorder cases and healthy siblings and parents. MPRAs have identified exonic variants that disrupt splicing (Rhine et al. 2022), 3′ UTR variants that modulate polyadenylation (Linder et al. 2022a), and 5′ UTR variants that impact translation and mRNA abundance (Plassmeyer et al. 2023) related to autism spectrum disease. Although further follow-up is required, the results from these assays provide a valuable resource for clinical practitioners and contribute to uncovering the genetic underpinnings of autism.

Even with MPRAs, it remains impossible to experimentally measure the effects of all possible variants, as the human genome supports billions of potential single-nucleotide variants and even more deletions and insertions. Models trained on MPRA data can bypass experimental throughput limits and bring variant analysis to the genome scale. Highlighting such scalability, Linder et al. (2022a) performed computational saturation mutagenesis of all human PASs, resulting in >40 million variant impact predictions and the identification of variants capable of disrupting polyadenylation.

Machine learning-guided sequence design for mRNA and gene therapy

Many therapeutic applications can benefit from rationally designed sequence elements. For example, engineered enhancers could increase gene therapy specificity and reduce side effects. Recent studies reported using machine learning-guided design to generate synthetic, cell type-specific enhancers (Gosai et al. 2023; De Almeida et al. 2024; Taskiran et al. 2024; Yin et al. 2024), often relying on MPRAs for model training or enhancer validation.

Similarly, synthetic UTRs for mRNA vaccines could increase protein expression compared with transcripts relying on human gene-derived sequence elements. Different design approaches using MPRAs and ML were used to generate 5′ UTRs for yeast and human mRNAs (Cuperus et al. 2017; Sample et al. 2019; Cao et al. 2021; Castillo-Hair et al. 2024), and synthetic UTRs were successful at driving high levels of protein production as desired for gene editing and other practical applications. However, 5′ UTRs could also be designed to target intermediate expression levels, demonstrating that the models comprehensively capture the underlying sequence–function relationship. Outside of potential gene and mRNA therapy applications, two complementary studies describe using MPRAs and ML to understand and engineer conditional translation control through toehold switches in bacteria (Angenent-Mari et al. 2020; Valeri et al. 2020).

To develop a conditional gene therapy using splicing, North et al. (2022) designed synthetic introns to be spliced in cancer cells with core splicing factor 3B1 (SF3B1) mutations but not in wild-type cells (North et al. 2022). Mutation-sensitive introns were identified from RNA sequencing data and low-throughput assays before characterization with a splicing MPRA. Synthetic introns were inserted into the herpes simplex virus–thymidine kinase (HSV-TK) system, such that functional HSV-TK would only be produced in cancer cells. Treatment of HSV-TK-expressing cells with the prodrug ganciclovir resulted in cytotoxic metabolite production and conditional cell death.

Given the importance of efficient CPA for generating stable transcripts, engineering PASs encouraging higher gene expression and targeted cleaving could have many applications in gene therapy. Linder et al. (2020) introduced novel sequence design algorithms and validated them by engineering polyadenylation signals in the 3′ UTR. PASs were designed to produce specific levels of 3′ end processing or to cleave a transcript at a defined distance from the core hexamer (Bogard et al. 2019; Linder et al. 2020).

Using MPRAs to characterize responses to drugs and perturbations

An intriguing application for MPRAs and ML is studying the *cis*-regulatory response to perturbations like drug treatments or varying expression of *trans*-regulators. Recent work used a cell-free polyadenylation MPRA and a CNN trained on these data to understand why some PASs are sensitive to treatment with the small molecule drug JTE-607 while others are not (Liu et al. 2023). Drug sensitivity was found to derive from a competition for binding to the polyadenylation machinery between the drug molecule and the mRNA sequence flanking the cleavage site. The model learned to accurately predict drug sensitivity of sequences not seen during training, and model interpretation identified features in the cleavage sequence conferring drug sensitivity.

An exon-skipping MPRA was used to map the 5′ SD sequence determinants of two splice-modifying small molecule drugs, risdiplam and branaplam, which were developed to treat spinal muscular atrophy by promoting the inclusion of SMN2 exon 7 (Ishigami et al. 2024). A biophysical model building on the MPRA data suggested a novel mechanism of drug–SD interactions and could explain the specificity differences between the two drugs. Additionally, two recent studies took advantage of MPRAs to characterize splicing changes in response to mutations in the SF3B1 (Gupta et al. 2019; North et al. 2022), highlighting the utility of MPRAs to map regulatory responses to changes in the *trans*-regulatory environment.

Challenges and future directions

Integration across processes and regions

What is next for MPRAs and models trained on them? One exciting direction for future MPRAs is learning to integrate information from different gene regions and processes through experiments and modeling, adding back some of the complexity removed in traditional MPRA designs. For example, it is increasingly clear that translation is coupled to mRNA stability (Wu and Bazzini 2023), with recent MPRA work finding that the presence of uORFs (Jia et al. 2020; Musaev et al. 2024)—but also features of the main ORF such as its length (Musaev et al. 2024)—impacts mRNA stability and not just translation. An MPRA combining flow-seq and RNA-seq to study the impact of transcript leader sequences on transcription efficiency found that splice donor sequences could increase transcript abundance, providing another example of coupling between multiple processes (Vlaming et al. 2022). One way that combined processes can be modeled is with multitask architectures with multiple outputs per input, such as those that can take in a single sequence and predict its translation efficiency and stability. Still, further work is necessary to comprehensively model and explain such observations.

Several MPRAs have already been designed to interrogate the interplay between multiple regulatory regions with the goal of developing transcript or gene-wide models (Leppek et al. 2022; O'Connell et al. 2023). A fundamental challenge for experiments aiming to vary multiple gene regions simultaneously is the explosive growth of possible combinations. For example, testing every core promoter with every putative enhancer or testing every 5′ UTR with every 3′ UTR in the human genome would require libraries of 10^8–10^{10} constructs. Still, this could be addressed if combinations are carefully selected. For instance, naturally co-occurring sequence motifs could be tested because they are more likely to exhibit nonlinear interactions than random motif combinations. Undoubtedly, as technology progresses, the availability of increased computing and sequencing capacity will aid these efforts. We also foresee continued development of modular models trained on large-scale region-specific MPRA data but fine-tuned with transcriptomic data or smaller customized MPRAs explicitly varying more than one functional gene region.

A second challenge intrinsic to MPRA design is that regulatory interactions beyond certain scales, approximately set by the length of available synthetic DNA, are not easily captured. These limitations mean that most splicing MPRAs focus on short exons, and enhancer or UTR MPRAs test fragments rather than full-length sequences. Long-read sequencing and gene synthesis methods can address some of these limitations but might still struggle with characterizing longer-range interactions, like those between enhancers and their cognate promoters. Genomically integrated MPRAs, possibly with perturbations made to nearby genomic regulatory elements, could help us better understand such interactions.

Finally, there are exciting possibilities for nucleotide LLMs and MPRAs. These nucleotide foundation models have a more stringent alphabet to work with than their protein counterparts, but promising work has shown that they have the potential to provide pretrained, multispecies starting points for modeling *cis*-regulatory processes (Consens et al. 2023; Dalla-Torre et al. 2023; Chu et al. 2024; Karollus et al. 2024; Nguyen et al. 2024). Importantly, such models can capture long-range dependencies in sequence space and could help bridge the gap between MPRA and genomic data modalities. We expect the development of nucleotide LLMs to continue, with specialized LLMs per regulatory region being a strong possibility given recent results (Tang and Koo 2024; Vilov and Heinig 2024).

Exploring additional modalities

Another future direction is applying MPRAs to explore novel regulatory mechanisms and contexts. For example, a recent MPRA was made to discern how 3′ UTR sequence controls subcellular mRNA localization in neuron-like cells (Mikl et al. 2022), and we expect to see more spatial MPRAs in the coming years. Future protein-sequencing approaches may enable direct, multiplexed protein measurements. Potentially soon, protein barcodes and nanopore sequencing might enable protein-level quantification (Cardozo et al. 2022), akin to using DNA and RNA barcodes in current MPRAs.

Chemical mRNA modifications such as N^6-methyladenosine (m^6A) or pseudouridine (Ψ) affect all aspects of the mRNA life cycle (Gilbert and Nachtergaele 2023; Delaunay et al. 2024). Given the importance of such modifications for native mRNA function and metabolism, it is likely that MPRAs will be performed to interrogate the *cis*-regulatory codes governing both their deposition and interpretation. Modified nucleotides, including Ψ and N1-methylpseudouridine (N1Ψ), are widely used to reduce unwanted innate immune activation of mRNA therapies and vaccines (Delaunay et al. 2024; Metkar et al. 2024). MPRAs are well suited to interrogate how such modifications might change stability, translation, or immune responses compared with "naked" IVT mRNA (Sample et al. 2019).

Similarly, we anticipate learning more about the impact of genomic position and chromatin state on gene expression and processing from MPRA experiments. Genomically integrated reporter systems are likely necessary for fully understanding cotranscriptional processing, whereas episomes or IVT RNA may be better suited for characterizing post-transcriptional regulation. Prior studies using integrated promoter and enhancer MPRAs suggest that physiological chromatinization impacts MPRA results while remaining consistent with the idea that episomal MPRAs accurately capture relative enhancer or promoter activities. Specifically, an MPRA testing the same enhancers on episomes and randomly integrated into the genome (Inoue et al. 2017) found that activities were well correlated between the two settings (Spearman $r = 0.79$). However, the activity range observed for integrated

reporters was wider, and reporter activity correlated better with genomic annotations for the enhancer sequences in their original genomic context. An MPRA comparing the same library of enhancers inserted at multiple genomic locations and on plasmids found that relative promoter strength was maintained across contexts, whereas absolute expression levels varied (Maricque et al. 2019). In complementary work, a few reporter constructs containing either an enhancer or a promoter were randomly integrated into thousands of genomic loci to comprehensively map the impact of genome position on CRE activity (Akhtar et al. 2013, 2014; Leemans et al. 2019). Mapping of insertion sites revealed large variations in expression levels linked to chromatin state and intrinsic promoter characteristics. In the future, we expect such approaches to be generalized to splicing and polyadenylation, allowing us to discover whether chromatin directly modulates these processes or does so indirectly through transcription regulation.

MPRAs in primary cells and tissues

Most MPRAs have been performed in a small number of cell lines. It is critical to answer to what extent *cis*-regulatory codes learned in one cell line can generalize to others and to cell types in the human body. Overall, a picture has emerged in which MPRAs in cell lines can effectively and quantitatively reveal aspects of shared, core regulatory codes but only capture some cell type-specific regulation.

Determining how additional data sources can augment MPRA measurements to allow predictions to generalize to new contexts may help solve this issue. One approach is to fine-tune models trained on high-quality MPRA data sets with other cell- or tissue-specific RNA-seq data sets. This approach was used successfully to predict the tissue-specific impact of variants on AS and APA (Bogard et al. 2019; Cheng et al. 2021; Linder et al. 2022a). Still, only relatively minor differences are observed at the tissue level, with more extensive variation likely occurring at the single-cell level. However, fine-tuning MPRA-trained models with single-cell RNA-seq data sets to make predictions for primary cell types remains an open challenge.

A complementary approach is performing MPRAs in primary cells and animal models, often with AAV or lentivirus delivery. Examples include enhancer MPRAs in mouse hepatocytes (Bravo González-Blas et al. 2024), primary cell types from the developing human cortex (Deng et al. 2024), human cortical organoids (Noack et al. 2023), lipopolysaccharide-activated B cells (Chaudhri et al. 2020), or mouse brains (Shen et al. 2016; Lambert et al. 2021; Noack et al. 2022). Nonenhancer examples include translation MPRAs performed with activated T cells (Castillo-Hair et al. 2024) and translation or stability MPRAs in developing zebrafish embryos (Rabani et al. 2017; Strayer et al. 2023; Reimão-Pinto et al. 2023). Such assays are sometimes combined with cell sorting to ensure that MPRA activity represents the cell type of interest.

Single-cell MPRA (scMPRA) technologies will enable us to resolve differences in gene regulation between cell types (Zhao et al. 2023b; Lalanne et al. 2024; Yin et al. 2024). In scMPRAs, native transcriptomics data can be used to group cells into cell types, whereas MPRA data from the same set of cells can resolve *cis*-regulatory grammar specific to each cell type. Although scMPRAs have only been used to study enhancers, adapting them to investigate other regulatory processes holds promise. Generalization of scMPRAs to splicing, polyadenylation, or mRNA abundance seems relatively straightforward, but developing stability time-course and translation assays may be more technically challenging.

Still, scMPRAs are not a panacea for resolving cell type-specific *cis*-regulatory codes. Although we expect to see more MPRAs in primary cells, tissues, and organs, delivering large, diverse libraries will remain challenging in many contexts. Moreover, single-cell readouts still face a problem of scale, as the diversity of the library that can be tested is likely inversely proportional to the diversity of cell types being interrogated, assuming a fixed sequencing budget and requirement to see each variant multiple times.

As such, a final critical long-term challenge is learning to predict how a *cis*-regulatory sequence will respond in a given cell type, even those inaccessible for MPRA experiments. Could we combine *cis*-regulatory MPRAs with libraries of *trans*-acting regulators such as transcription factors (Ng et al. 2021; Joung et al. 2023) to learn specific regulatory relationships that can be generalized to novel cell types, given knowledge of the *trans*-regulatory environment in that cell type? Given these exciting opportunities, MPRAs and models trained on them will continue to grow and play an increasingly important role in future studies of gene regulation in health and disease.

Competing interest statement

G.S. is a cofounder of Parse Biosciences and an advisor to Deep Genomics and Sanofi.

Acknowledgments

Y.S. acknowledges support through National Institutes of Health (NIH) Awards R35GM149294 and R01AI155962. G.S. acknowledges support through National Science Foundation Award 2021552 and NIH Award R01GM149631.

References

Adamson SI, Zhan L, Graveley BR. 2018. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol* **19:** 71. doi:10.1186/s13059-018-1437-x

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **76:** 7–20. doi:10.1002/0471142905.hg0720s76

Agarwal V, Kelley DR. 2022. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol* **23:** 245. doi:10.1186/s13059-022-02811-x

Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LFA, van Lohuizen M, van Steensel B.

2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154:** 914–927. doi:10.1016/j.cell.2013.07.018

Akhtar W, Pindyurin AV, De Jong J, Pagie L, Ten Hoeve J, Berns A, Wessels LFA, Van Steensel B, Van Lohuizen M. 2014. Using TRIP for genome-wide position effect analysis in cultured cells. *Nat Protoc* **9:** 1255–1281. doi:10.1038/nprot.2014.072

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33:** 831–838. doi:10.1038/nbt.3300

Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, Hart M, Abu-Halima M, Grässer FA, Lenhof H-P, et al. 2019. An estimate of the total number of true human miRNAs. *Nucleic Acids Res* **47:** 3353–3364. doi:10.1093/nar/gkz097

Angenent-Mari NM, Garruss AS, Soenksen LR, Church G, Collins JJ. 2020. A deep learning approach to programmable RNA switches. *Nat Commun* **11:** 5057. doi:10.1038/s41467-020-18677-1

Angermueller C, Lee HJ, Reik W, Stegle O. 2017. Deepcpg: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18:** 67. doi:10.1186/s13059-017-1189-z

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077. doi:10.1126/science.1232542

Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. 2019. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol* **20:** 183. doi:10.1186/s13059-019-1787-z

Avdeyev P, Shi C, Tan Y, Dudnyk K, Zhou J. 2023. Dirichlet diffusion score model for biological sequence generation. arXiv doi:10.48550/arXiv.2305.10699

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53:** 354–366. doi:10.1038/s41588-021-00782-6

Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. 2019. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* **176:** 549–563.e23. doi:10.1016/j.cell.2018.12.010

Baeza-Centurion P, Miñana B, Valcárcel J, Lehner B. 2020. Mutations primarily alter the inclusion of alternatively spliced exons. *eLife* **9:** e59959. doi:10.7554/eLife.59959

Barreau C. 2005. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res* **33:** 7138–7150. doi:10.1093/nar/gki1012

Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. 2021. Low-N protein engineering with data-efficient deep learning. *Nat Methods* **18:** 389–396. doi:10.1038/s41592-021-01100-y

Blair JD, Hockemeyer D, Doudna JA, Bateup HS, Floor SN. 2017. Widespread translational remodeling during human neuronal differentiation. *Cell Rep* **21:** 2005–2016. doi:10.1016/j.celrep.2017.10.095

Bogard N, Linder J, Rosenberg AB, Seelig G. 2019. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178:** 91–106.e23. doi:10.1016/j.cell.2019.04.046

Bomba L, Walter K, Soranzo N. 2017. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* **18:** 77. doi:10.1186/s13059-017-1212-4

Braun S, Enculescu M, Setty ST, Cortés-López M, De Almeida BP, Reymond Sutandy FX, Schulz L, Busch A, Seiler M, Ebersberger S, et al. 2018. Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat Commun* **9:** 3315. doi:10.1038/s41467-018-05748-7

Bravo González-Blas C, Matetovici I, Hillen H, Taskiran II, Vandepoel R, Christiaens V, Sansores-García L, Verboven E, Hulselmans G, Poovathingal S, et al. 2024. Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation. *Nat Cell Biol* **26:** 153–167. doi:10.1038/s41556-023-01316-4

Brookes DH, Listgarten J. 2018. Design by adaptive sampling. arXiv doi:10.48550/ARXIV.1810.03714

Cambray G, Guimaraes JC, Arkin AP. 2018. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. *Nat Biotechnol* **36:** 1005–1015. doi:10.1038/nbt.4238

Cao J, Novoa EM, Zhang Z, Chen WCW, Liu D, Choi GCG, Wong ASL, Wehrspaun C, Kellis M, Lu TK. 2021. High-throughput 5′ UTR engineering for enhanced protein production in non-viral gene therapies. *Nat Commun* **12:** 4138. doi:10.1038/s41467-021-24436-7

Cardozo N, Zhang K, Doroschak K, Nguyen A, Siddiqui Z, Bogard N, Strauss K, Ceze L, Nivala J. 2022. Multiplexed direct detection of barcoded protein reporters on a nanopore array. *Nat Biotechnol* **40:** 42–46. doi:10.1038/s41587-021-01002-6

Carter B, Bileschi M, Smith J, Sanderson T, Bryant D, Belanger D, Colwell LJ. 2020. Critiquing protein family classification models using sufficient input subsets. *J Comput Biol* **27:** 1219–1231. doi:10.1089/cmb.2019.0339

Castillo-Hair S, Fedak S, Wang B, Linder J, Havens K, Certo M, Seelig G. 2024. Optimizing 5′UTRs for mRNA-delivered gene editing using deep learning. *Nat Commun* **15:** 5284. doi:10.1038/s41467-024-49508-2

Chan S, Choi E, Shi Y. 2011. Pre-mRNA 3′-end processing complex assembly and function. *Wiley Interdiscip Rev RNA* **2:** 321–335. doi:10.1002/wrna.54

Chaudhri VK, Dienger-Stambaugh K, Wu Z, Shrestha M, Singh H. 2020. Charting the *cis*-regulome of activated B cells by coupling structural and functional genomics. *Nat Immunol* **21:** 210–220. doi:10.1038/s41590-019-0565-0

Chen I-T, Chasin LA. 1993. Direct selection for mutations affecting specific splice sites in a hamster dihydrofolate reductase minigene. *Mol Cell Biol* **13:** 289–300. doi:10.1128/mcb.13.1.289-300.1993

Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, Zong L, Hong L, Xiao J, Shen T, et al. 2022. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. bioRxiv doi:10.1101/2022.08.06.503062

Chen K, Zhou Y, Ding M, Wang Y, Ren Z, .Yang Y. 2024. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. *Brief Bioinform* **25:** bbae163. doi:10.1093/bib/bbae163

Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Avsec Ž, Gagneur J. 2019. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* **20:** 48. doi:10.1186/s13059-019-1653-z

Cheng J, Çelik MH, Kundaje A, Gagneur J. 2021. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biol* **22:** 94. doi:10.1186/s13059-021-02273-7

Chiang H-L, Chen Y-T, Su J-Y, Lin H-N, Yu C-HA, Hung Y-J, Wang Y-L, Huang Y-T, Lin C-L. 2022. Mechanism and modeling of human disease-associated near-exon intronic variants that perturb RNA splicing. *Nat Struct Mol Biol* **29:** 1043–1055. doi:10.1038/s41594-022-00844-1

Chong R, Insigne KD, Yao D, Burghard CP, Wang J, Hsiao Y-HE, Jones EM, Goodman DB, Xiao X, Kosuri S. 2019. A

multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Mol Cell* **73:** 183–194.e8. doi:10.1016/j.molcel.2018.10.037

Chu Y, Yu D, Li Y, Huang K, Shen Y, Cong L, Zhang J, Wang M. 2024. A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nat Mach Intell* **6:** 449–460. doi:10.1038/s42256-024-00823-9

Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, Theis FJ, Moses A, Wang B. 2023. To transformers and beyond: large language models for the genome. arXiv doi:10.48550/ARXIV.2311.07621

Cortés-López M, Schulz L, Enculescu M, Paret C, Spiekermann B, Quesnel-Vallières M, Torres-Diz M, Unic S, Busch A, Orekhova A, et al. 2022. High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling CD19 splicing and CART-19 therapy resistance. *Nat Commun* **13:** 5570. doi:10.1038/s41467-022-31818-y

Costello Z, Martin HG. 2019. How to hallucinate functional proteins. arXiv doi:10.48550/ARXIV.1903.00458

Cottrell KA, Chaudhari HG, Cohen BA, Djuranovic S. 2018. PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs. *Nat Commun* **9:** 301. doi:10.1038/s41467-017-02745-0

Culler SJ, Hoff KG, Voelker RB, Andrew Berglund J, Smolke CD. 2010. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Res* **38:** 5152–5165. doi:10.1093/nar/gkq248

Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, Seelig G. 2017. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res* **27:** 2015–2024. doi:10.1101/gr.224964.117

Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Carranza NL, Grzywaczewski AH, Oteri F, Dallago C, Trop E, de Almeida BP, Sirelkhatim H, et al. 2023. The nucleotide transformer: building and evaluating robust foundation models for human genomics. bioRxiv doi:10.1101/2023.01.11.523679

DaSilva LF, Senan S, Patel ZM, Reddy AJ, Gabbita S, Nussbaum Z, Córdova CMV, Wenteler A, Weber N, Tunjic TM, et al. 2024. DNA-diffusion: leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements. bioRxiv doi:10.1101/2024.02.01.578352

De Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54:** 613–624. doi:10.1038/s41588-022-01048-5

De Almeida BP, Schaub C, Pagani M, Secchia S, Furlong EEM, Stark A. 2024. Targeted design of synthetic enhancers for selected tissues in the *Drosophila* embryo. *Nature* **626:** 207–211. doi:10.1038/s41586-023-06905-9

De Boer CG, Taipale J. 2024. Hold out the genome: a roadmap to solving the *cis*-regulatory code. *Nature* **625:** 41–50. doi:10.1038/s41586-023-06661-w

De Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38:** 56–65. doi:10.1038/s41587-019-0315-8

Delaunay S, Helm M, Frye M. 2024. RNA modifications in physiology and disease: towards clinical applications. *Nat Rev Genet* **25:** 104–122. doi:10.1038/s41576-023-00645-2

Deng C, Whalen S, Steyert M, Ziffra R, Przytycki PF, Inoue F, Pereira DA, Capauto D, Norton S, Vaccarino FM, et al. 2024. Massively parallel characterization of regulatory elements in the developing human cortex. *Science* **384:** eadh0559. doi:10.1126/science.adh0559

Derti A, Garrett-Engele P, MacIsaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22:** 1173–1183. doi:10.1101/gr.132563.111

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci* **110:** E2792–E2801. doi:10.1073/pnas.1222534110

Ellington AD, Szostak JW. 1990. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346:** 818–822. doi:10.1038/346818a0

Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nat Rev Gene* **20:** 389–403. doi:10.1038/s41576-019-0122-6

Faure AJ, Schmiedel JM, Baeza-Centurion P, Lehner B. 2020. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol* **21:** 207. doi:10.1186/s13059-020-02091-3

Ferreira JP, Overton KW, Wang CL. 2013. Tuning gene expression with synthetic upstream open Reading frames. *Proc Natl Acad Sci* **110:** 11284–11289. doi:10.1073/pnas.1305590110

Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. 2014. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513:** 120–123. doi:10.1038/nature13695

Floor SN, Doudna JA. 2016. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5:** e10921. doi:10.7554/eLife.10921

Friedman RZ, Ramu A, Lichtarge S, Myers CA, Granas DM, Gause M, Corbo JC, Cohen BA, White MA. 2023. Active learning of enhancer and silencer regulatory grammar in photoreceptors. bioRxiv doi:10.1101/2023.08.21.554146

Fu T, Amoah K, Chan TW, Bahn JH, Lee J-H, Terrazas S, Chong R, Kosuri S, Xiao X. 2024. Massively parallel screen uncovers many rare 3′ UTR variants regulating mRNA abundance of cancer driver genes. *Nat Commun* **15:** 3335. doi:10.1038/s41467-024-46795-7

Gallego Romero I, Lea AJ. 2023. Leveraging massively parallel reporter assays for evolutionary questions. *Genome Biol* **24:** 26. doi:10.1186/s13059-023-02856-6

Gergics P, Smith C, Bando H, Jorge AAL, Rockstroh-Lippold D, Vishnopolska SA, Castinetti F, Maksutova M, Carvalho LRS, Hoppmann J, et al. 2021. High-throughput splicing assays identify missense and silent splice-disruptive POU1F1 variants underlying pituitary hormone deficiency. *Am J Hum Genet* **108:** 1526–1539. doi:10.1016/j.ajhg.2021.06.013

Gilbert WV, Nachtergaele S. 2023. mRNA regulation by RNA modifications. *Annu Rev Biochem* **92:** 175–198. doi:10.1146/annurev-biochem-052521-035949

Glidden DT, Buerer JL, Saueressig CF, Fairbrother WG. 2021. Hotspot exons are common targets of splicing perturbations. *Nat Commun* **12:** 2756. doi:10.1038/s41467-021-22780-2

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative adversarial networks. arXiv. doi:10.48550/arXiv.1406.2661

Gordon MG, Inoue F, Martin B, Schubach M, Agarwal V, Whalen S, Feng S, Zhao J, Ashuach T, Ziffra R, et al. 2020. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat Protoc* **15:** 2387–2412. doi:10.1038/s41596-020-0333-5

Gosai S, Castro R, Fuentes N, Butts J, Kales S, Noche R, Mouri K, Sabeti P, Reilly S, Tewhey R. 2023. Machine-guided design of

synthetic cell type-specific *cis*-regulatory elements. bioRxiv doi:10.1101/2023.08.08.552077

Graveley BR, Fleming ES, Gilmartin GM. 1996. RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. *Mol Cell Biol* **16:** 4942–4951. doi:10.1128/MCB.16.9.4942

Greenside P, Shimko T, Fordyce P, Kundaje A. 2018. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* **34:** i629–i637. doi:10.1093/bioinformatics/bty575

Griesemer D, Xue JR, Reilly SK, Ulirsch JC, Kukreja K, Davis JR, Kanai M, Yang DK, Butts JC, Guney MH, et al. 2021. Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184:** 5247–5260.e19. doi:10.1016/j.cell.2021.08.025

Gruber AJ, Zavolan M. 2019. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* **20:** 599–614. doi:10.1038/s41576-019-0145-z

Gruber AJ, Schmidt R, Gruber AR, Martin G, Ghosh S, Belmadani M, Keller W, Zavolan M. 2016. A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res* **26:** 1145–1159. doi:10.1101/gr.202432.115

Gupta A, Zou J. 2019. Feedback GAN for DNA optimizes protein functions. *Nat Mach Intell* **1:** 105–111. doi:10.1038/s42256-019-0017-4

Gupta AK, Murthy T, Paul KV, Ramirez O, Fisher JB, Rao S, Rosenberg AB, Seelig G, Minella AC, Pillai MM. 2019. Degenerate minigene library analysis enables identification of altered branch point utilization by mutant splicing factor 3B1 (SF3B1). *Nucleic Acids Res* **47:** 970–980. doi:10.1093/nar/gky1161

Gupta K, Yang C, McCue K, Bastani O, Sharp PA, Burge CB, Solar-Lezama A. 2024. Improved modeling of RNA-binding protein motifs in an interpretable neural model of RNA splicing. *Genome Biol* **25:** 23. doi:10.1186/s13059-023-03162-x

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks LA, Welbl J, Clark A, et al. 2022. Training compute-optimal large language models. arXiv doi:10.48550/ARXIV.2203.15556

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods* **10:** 133–139. doi:10.1038/nmeth.2288

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27:** 38–52. doi:10.1101/gr.212092.116

Ipe J, Collins KS, Hao Y, Gao H, Bhatia P, Gaedigk A, Liu Y, Skaar TC. 2018. PASSPORT-seq: a novel high-throughput bioassay to functionally test polymorphisms in micro-RNA target sites. *Front Genet* **9:** 219. doi:10.3389/fgene.2018.00219

Ishigami Y, Wong MS, Martí-Gómez C, Ayaz A, Kooshkbaghi M, Hanson SM, McCandlish DM, Krainer AR, Kinney JB. 2024. Specificity, synergy, and mechanisms of splice-modifying drugs. *Nat Commun* **15:** 1880. doi:10.1038/s41467-024-46090-5

Ji Y, Zhou Z, Liu H, Davuluri RV. 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37:** 2112–2120. doi:10.1093/bioinformatics/btab083

Jia L, Mao Y, Ji Q, Dersh D, Yewdell JW, Qian S-B. 2020. Decoding mRNA translatability and stability from the 5′ UTR. *Nat Struct Mol Biol* **27:** 814–821. doi:10.1038/s41594-020-0465-x

Joung J, Ma S, Tay T, Geiger-Schuller KR, Kirchgatterer PC, Verdine VK, Guo B, Arias-Garcia MA, Allen WE, Singh A, et al. 2023. A transcription factor atlas of directed differentiation. *Cell* **186:** 209–229.e26. doi:10.1016/j.cell.2022.11.026

Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. 2016. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nat Commun* **7:** 11558. doi:10.1038/ncomms11558

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581:** 434–443. doi:10.1038/s41586-020-2308-7

Karollus A, Avsec Ž, Gagneur J. 2021. Predicting mean ribosome load for 5′UTR of any length using deep learning. *PLoS Comput Biol* **17:** e1008982. doi:10.1371/journal.pcbi.1008982

Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J. 2024. Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol* **25:** 83. doi:10.1186/s13059-024-03221-x

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* **21:** 1360–1374. doi:10.1101/gr.119628.110

Ke S, Anquetil V, Zamalloa JR, Maity A, Yang A, Arias MA, Kalachikov S, Russo JJ, Ju J, Chasin LA. 2018. Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res* **28:** 11–24. doi:10.1101/gr.219683.116

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23:** 800–811. doi:10.1101/gr.144899.112

Killoran N, Lee LJ, Delong A, Duvenaud D, Frey BJ. 2017. Generating and designing DNA with deep generative models. arXiv doi:10.48550/ARXIV.1712.06148

Kim I, Miller CR, Young DL, Fields S. 2013. High-throughput analysis of in vivo protein stability. *Mol Cell Proteomics* **12:** 3370–3378. doi:10.1074/mcp.O113.031708

Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. *arXiv* doi:10.48550/ARXIV.1312.6114

Kinney JB, McCandlish DM. 2019. Massively parallel assays and quantitative sequence–function relationships. *Annu Rev Genomics Hum Genet* **20:** 99–127. doi:10.1146/annurev-genom-083118-014845

Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci* **107:** 9158–9163. doi:10.1073/pnas.1004290107

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46:** 310–315. doi:10.1038/ng.2892

Kleftogiannis D, Kalnis P, Bajic VB. 2015. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res* **43:** e6. doi:10.1093/nar/gku1058

Klie A, Laub D, Talwar JV, Stites H, Jores T, Solvason JJ, Farley EK, Carter H. 2023. Predictive analyses of regulatory sequences with EUGENe. *Nat Comput Sci* **3:** 946–956. doi:10.1038/s43588-023-00544-w

Koo PK, Eddy SR. 2019. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput Biol* **15:** e1007560. doi:10.1371/journal.pcbi.1007560

Koo PK, Ploenzke M. 2021. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nat Mach Intell* **3:** 258–266. doi:10.1038/s42256-020-00291-x

Koo PK, Qian S, Kaplun G, Volf V, Kalimeris D. 2019. Robust neural networks are more interpretable for genomics. bioRxiv doi:10.1101/657437

Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM. 2013. Composability of regulatory sequences controlling transcription and translation in *escherichia coli*. *Proc Natl Acad Sci* **110:** 14024–14029. doi:10.1073/pnas.1301301110

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc Natl Acad Sci* **109:** 19498–19503. doi:10.1073/pnas.1210678109

Lal A, Garfield D, Biancalani T, Eraslan G. 2024. Designing realistic regulatory DNA with autoregressive language models. *Genome Res* doi:10.1101/gr.279142.124

Lalanne J-B, Regalado SG, Domcke S, Calderon D, Martin BK, Li X, Li T, Li T, Suiter CC, Lee C, et al. 2024. Multiplex profiling of developmental *cis*-regulatory elements with quantitative single-cell expression reporters. *Nat Methods* **21:** 983–993. doi:10.1038/s41592-024-02260-3

Lambert JT, Su-Feher L, Cichewicz K, Warren TL, Zdilar I, Wang Y, Lim KJ, Haigh JL, Morse SJ, Canales CP, et al. 2021. Parallel functional testing identifies enhancers active in early postnatal mouse brain. *eLife* **10:** e69479. doi:10.7554/eLife.69479

Lanchantin J, Singh R, Lin Z, Qi Y. 2016. Deep motif: visualizing genomic sequence classifications. arXiv doi:10.48550/ARXIV.1605.01133

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46:** D1062–D1067. doi:10.1093/nar/gkx1153

Leemans C, Van Der Zwalm MCH, Brueckner L, Comoglio F, Van Schaik T, Pagie L, Van Arensbergen J, Van Steensel B. 2019. Promoter-intrinsic and local chromatin features determine gene repression in LADs. *Cell* **177:** 852–864.e14. doi:10.1016/j.cell.2019.03.009

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536:** 285–291. doi:10.1038/nature19057

Leppek K, Das R, Barna M. 2018. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol* **19:** 158–174. doi:10.1038/nrm.2017.103

Leppek K, Byeon GW, Kladwang W, Wayment-Steele HK, Kerr CH, Xu AF, Kim DS, Topkar VV, Choe C, Rothschild D, et al. 2022. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat Commun* **13:** 1536. doi:10.1038/s41467-022-28776-w

Letiagina AE, Omelina ES, Ivankin AV, Pindyurin AV. 2021. MPRAdecoder: processing of the raw MPRA data with a priori unknown sequences of the region of interest and associated barcodes. *Front Genet* **12:** 618189. doi:10.3389/fgene.2021.618189

Li Z, Li Y, Zhang B, Li Y, Long Y, Zhou J, Zou X, Zhang M, Hu Y, Chen W, et al. 2022. DeeReCT-APA: prediction of alternative polyadenylation site usage through deep learning. *Genomics Proteomics Bioinformatics* **20:** 483–495. doi:10.1016/j.gpb.2020.05.004

Liachko I, Youngblood RA, Keich U, Dunham MJ. 2013. High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. *Genome Res* **23:** 698–704. doi:10.1101/gr.144659.112

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27:** 2380–2396. doi:10.1101/gad.229328.113

Liao SE, Sudarshan M, Regev O. 2023. Deciphering RNA splicing logic with interpretable machine learning. *Proc Natl Acadf Sci* **120:** e2221165120. doi:10.1073/pnas.2221165120

Lim Y, Arora S, Schuster SL, Corey L, Fitzgibbon M, Wladyka CL, Wu X, Coleman IM, Delrow JJ, Corey E, et al. 2021. Multiplexed functional genomic analysis of 5′ untranslated region mutations across the spectrum of prostate cancer. *Nat Commun* **12:** 4217. doi:10.1038/s41467-021-24445-6

Lin Y, May GE, Kready H, Nazzaro L, Mao M, Spealman P, Creeger Y, Joel McManus C. 2019. Impacts of uORF codon identity and position on translation regulation. *Nucleic Acids Res* **47:** 9358–9367. doi:10.1093/nar/gkz681

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379:** 1123–1130. doi:10.1126/science.ade2574

Linder J, Seelig G. 2021. Fast activation maximization for molecular sequence design. *BMC Bioinformatics* **22:** 510. doi:10.1186/s12859-021-04437-5

Linder J, Bogard N, Rosenberg AB, Seelig G. 2020. A generative neural network for maximizing fitness and diversity of synthetic DNA and protein sequences. *Cell Syst* **11:** 49–62.e16. doi:10.1016/j.cels.2020.05.007

Linder J, Koplik SE, Kundaje A, Seelig G. 2022a. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol* **23:** 232. doi:10.1186/s13059-022-02799-4

Linder J, La Fleur A, Chen Z, Ljubetič A, Baker D, Kannan S, Seelig G. 2022b. Interpreting neural networks for biological sequences by learning stochastic masks. *Nat Mach Intell* **4:** 41–54. doi:10.1038/s42256-021-00428-6

Litterman AJ, Kageyama R, Le Tonqueze O, Zhao W, Gagnon JD, Goodarzi H, Erle DJ, Mark Ansel K. 2019. A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res* **29:** 896–906. doi:10.1101/gr.242552.118

Liu L, Yu AM, Wang X, Soles LV, Teng X, Chen Y, Yoon Y, Sarkan KSK, Valdez MC, Linder J, et al. 2023. The anticancer compound JTE-607 reveals hidden sequence specificity of the mRNA 3′ processing machinery. *Nat Struct Mol Biol* **30:** 1947–1957. doi:10.1038/s41594-023-01161-x

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2:** 56–67. doi:10.1038/s42256-019-0138-9

Marasco LE, Kornblihtt AR. 2023. The physiology of alternative splicing. *Nat Rev Mol Cell Biol* **24:** 242–254. doi:10.1038/s41580-022-00545-z

Maricque BB, Chaudhari HG, Cohen BA. 2019. A massively parallel reporter assay dissects the influence of chromatin structure on *cis*-regulatory activity. *Nat Biotechnol* **37:** 90–95. doi:10.1038/nbt.4285

May GE, Akirtava C, Agar-Johnson M, Micic J, Woolford J, McManus J. 2023. Unraveling the influences of sequence and position on yeast uORF activity using massively parallel reporter systems and machine learning. *eLife* **12:** e69611. doi:10.7554/eLife.69611

Mayr C. 2017. Regulation by 3′-untranslated regions. *Annu Rev Genet* **51:** 171–194. doi:10.1146/annurev-genet-120116-024704

McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, Bartel DP. 2019. The biochemical basis of microRNA targeting efficacy. *Science* **366:** eaav1741. doi:10.1126/science.aav1741

McGillivray P, Ault R, Pawashe M, Kitchen R, Balasubramanian S, Gerstein M. 2018. A comprehensive catalog of predicted functional upstream open Reading frames in humans. *Nucleic Acids Res* **46:** 3326–3338. doi:10.1093/nar/gky188

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30:** 271–277. doi:10.1038/nbt.2137

Metkar M, Pepin CS, Moore MJ. 2024. Tailor made: the art of therapeutic mRNA design. *Nat Rev Drug Discov* **23:** 67–83. doi:10.1038/s41573-023-00827-x

Mikl M, Hamburg A, Pilpel Y, Segal E. 2019. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat Commun* **10:** 4572. doi:10.1038/s41467-019-12642-3

Mikl M, Eletto D, Nijim M, Lee M, Lafzi A, Mhamedi F, David O, Sain SB, Handler K, Moor AE. 2022. A massively parallel reporter assay reveals focused and broadly encoded RNA localization signals in neurons. *Nucleic Acids Res* **50:** 10643–10664. doi:10.1093/nar/gkac806

Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V, Makhzami S, Seltenhammer M, Karras P, et al. 2020. Cross-species analysis of enhancer logic using deep learning. *Genome Res* **30:** 1815–1834. doi:10.1101/gr.260844.120

Mitschka S, Mayr C. 2022. Context-specific regulation and function of mRNA alternative polyadenylation. *Nat Rev Mol Cell Biol* **23:** 779–796. doi:10.1038/s41580-022-00507-5

Mogno I, Kwasnieski JC, Cohen BA. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* **23:** 1908–1915. doi:10.1101/gr.157891.113

Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. 2019. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One* **14:** e0218073. doi:10.1371/journal.pone.0218073

Musaev D, Abdelmessih M, Vejnar CE, Yartseva V, Weiss LA, Strayer EC, Takacs CM, Giraldez AJ. 2024. UPF1 regulates mRNA stability by sensing poorly translated coding sequences. *Cell Rep* **43:** 114074. doi:10.1016/j.celrep.2024.114074

Myint L, Avramopoulos DG, Goff LA, Hansen KD. 2019. Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* **20:** 209. doi:10.1186/s12864-019-5556-x

Ng AHM, Khoshakhlagh P, Rojo Arias JE, Pasquini G, Wang K, Swiersy A, Shipman SL, Appleton E, Kiaee K, Kohman RE, et al. 2021. A comprehensive library of human transcription factors for cell fate engineering. *Nat Biotechnol* **39:** 510–519. doi:10.1038/s41587-020-0742-6

Nguyen E, Poli M, Durrant MG, Thomas AW, Kang B, Sullivan J, Ng MY, Lewis A, Patel A, Lou A, et al. 2024. Sequence modeling and design from molecular to genome scale with Evo. bioRxiv doi:10.1101/2024.02.27.582234

Niederer RO, Rojas-Duran MF, Zinshteyn B, Gilbert WV. 2022. Direct analysis of ribosome targeting illuminates thousandfold regulation of translation initiation. *Cell Syst* **13:** 256–264.e3. doi:10.1016/j.cels.2021.12.002

Nikolados E-M, Wongprommoon A, Aodha OM, Cambray G, Oyarzún DA. 2022. Accuracy and data efficiency in deep learning models of protein expression. *Nat Commun* **13:** 7755. doi:10.1038/s41467-022-34902-5

Noack F, Vangelisti S, Raffl G, Carido M, Diwakar J, Chong F, Bonev B. 2022. Multimodal profiling of the transcriptional regulatory landscape of the developing mouse cortex identifies Neurog2 as a key epigenome remodeler. *Nat Neurosci* **25:** 154–167. doi:10.1038/s41593-021-01002-4

Noack F, Vangelisti S, Ditzer N, Chong F, Albert M, Bonev B. 2023. Joint epigenome profiling reveals cell-type-specific gene regulatory programmes in human cortical organoids. *Nat Cell Biol* **25:** 1873–1883. doi:10.1038/s41556-023-01296-5

Noderer WL, Flockhart RJ, Bhaduri A, Diaz De Arce AJ, Zhang J, Khavari PA, Wang CL. 2014. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol Syst Biol* **10:** 748. doi:10.15252/msb.20145136

Norn C, Wicky BIM, Juergens D, Liu S, Kim D, Tischer D, Koepnick B, Anishchenko I, Baker D, Ovchinnikov S, et al. 2021. Protein sequence design by conformational landscape optimization. *Proc Natl Acad Sci* **118:** e2017228118. doi:10.1073/pnas.2017228118

North K, Benbarche S, Liu B, Pangallo J, Chen S, Stahl M, Bewersdorf JP, Stanley RF, Erickson C, Cho H, et al. 2022. Synthetic introns enable splicing factor mutation-dependent targeting of cancer cells. *Nat Biotechnol* **40:** 1103–1113. doi:10.1038/s41587-022-01224-2

Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. 2023. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet* **24:** 125–137. doi:10.1038/s41576-022-00532-2

O'Connell RW, Rai K, Piepergerdes TC, Wang Y, Samra KD, Wilson JA, Lin S, Zhang TH, Ramos EM, Sun A, et al. 2023. Ultra-high throughput mapping of genetic design space. bioRxiv doi:10.1101/2023.03.16.532704

Oikonomou P, Goodarzi H, Tavazoie S. 2014. Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. *Cell Rep* **7:** 281–292. doi:10.1016/j.celrep.2014.03.001

Oliphant AR, Struhl K. 1989. An efficient method for generating proteins with altered enzymatic properties: application to β-lactamase. *Proc Natl Acad Sci* **86:** 9094–9098. doi:10.1073/pnas.86.23.9094

Paggi J, Lamb A, Tian K, Hsu I, Cedoz P-L, Kawthekar P. 2017. Predicting transcriptional regulatory activities with deep convolutional networks. bioRxiv doi:10.1101/099879

Park M, Singh S, Khan SR, Abrar MA, Grisanti F, Rahman MS, Samee MAH. 2022. Multinomial convolutions for joint modeling of regulatory motifs and sequence activity readouts. *Genes* **13:** 1614. doi:10.3390/genes13091614

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27:** 1173–1175. doi:10.1038/nbt.1589

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30:** 265–270. doi:10.1038/nbt.2136

Penzar D, Nogina D, Noskova E, Zinkevich A, Meshcheryakov G, Lando A, Rafi AM, De Boer C, Kulakovskiy IV. 2023. Legnet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics* **39:** btad457. doi:10.1093/bioinformatics/btad457

Perchlik M, Sasse A, Mostafavi S, Fields S, Cuperus JT. 2024. Impact on splicing in *Saccharomyces cerevisiae* of random 50-base sequences inserted into an intron. *RNA* **30:** 52–67. doi:10.1261/rna.079752.123

Peterman N, Levine E. 2016. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17:** 206. doi:10.1186/s12864-016-2533-5

Plassmeyer SP, Florian CP, Kasper MJ, Chase R, Mueller S, Liu Y, White KM, Jungers CF, Djuranovic SP, Djuranovic S, et al. 2023. A massively parallel screen of 5′UTR mutations identifies variants impacting translation and protein production in neurodevelopmental disorder genes. medRxiv doi:10.1101/2023.11.02.23297961

Quang D, Xie X. 2016. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* **44:** e107. doi:10.1093/nar/gkw226

Rabani M, Pieper L, Chew G-L, Schier AF. 2017. A massively parallel reporter assay of 3′ UTR sequences identifies in vivo rules for mRNA degradation. *Mol Cell* **68:** 1083–1094.e5. doi:10.1016/j.molcel.2017.11.014

Reimão-Pinto MM, Castillo-Hair SM, Seelig G, Schier AF. 2023. The regulatory landscape of 5′ UTRs in translational control during zebrafish embryogenesis. bioRxiv doi:10.1101/2023.11.23.568470

Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, Poviloniene S, Laurynenas A, Viknander S, Abuajwa W, et al. 2021. Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell* **3:** 324–333. doi:10.1038/s42256-021-00310-5

Rhine CL, Neil C, Wang J, Maguire S, Buerer L, Salomon M, Meremikwu IC, Kim J, Strande NT, Fairbrother WG. 2022. Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of autism genes. *PLoS Genet* **18:** e1009884. doi:10.1371/journal.pgen.1009884

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* **118:** e2016239118. doi:10.1073/pnas.2016239118

Rogalska ME, Vivori C, Valcárcel J. 2023. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat Rev Genet* **24:** 251–269. doi:10.1038/s41576-022-00556-8

Rosati D, Palmieri M, Brunelli G, Morrione A, Iannelli F, Frullanti E, Giordano A. 2024. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: a review. *Comput Struct Biotechnol J* **23:** 1154–1168. doi:10.1016/j.csbj.2024.02.018

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163:** 698–711. doi:10.1016/j.cell.2015.09.054

Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM. 2017. A statistical framework for analyzing deep mutational scanning data. *Genome Biol* **18:** 150. doi:10.1186/s13059-017-1272-5

Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, Seelig G. 2019. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* **37:** 803–809. doi:10.1038/s41587-019-0164-5

Sarkar A, Tang Z, Zhao C, Koo P. 2024. Designing DNA with tunable regulatory activity using discrete diffusion. bioRxiv doi:10.1101/2024.05.23.595630

Savinov A, Brandsen BM, Angell BE, Cuperus JT, Fields S. 2021. Effects of sequence motifs in the yeast 3′ untranslated region determined from massively parallel assays of random sequences. *Genome Biol* **22:** 293. doi:10.1186/s13059-021-02509-6

Schirman D, Yakhini Z, Pilpel Y, Dahan O. 2021. A broad analysis of splicing regulation in yeast using a large library of synthetic introns. *PLoS Genet* **17:** e1009805. doi:10.1371/journal.pgen.1009805

Schuster SL, Arora S, Wladyka CL, Itagi P, Corey L, Young D, Stackhouse BL, Kollath L, Wu QV, Corey E, et al. 2023. Multi-level functional genomics reveals molecular and cellular oncogenicity of patient-based 3′ untranslated region mutations. *Cell Rep* **42:** 112840. doi:10.1016/j.celrep.2023.112840

Seitz EE, McCandlish DM, Kinney JB, Koo PK. 2024. Interpreting *cis*-regulatory mechanisms from genomic deep neural networks using surrogate models. *Nat Mach Intell* **6:** 701–713. doi:10.1038/s42256-024-00851-5

Seo JJ, Jung S-J, Yang J, Choi D-E, Narry Kim V. 2023. Functional viromic screens uncover regulatory RNA elements. *Cell* **186:** 3291–3306.e21. doi:10.1016/j.cell.2023.06.007

Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, Segal E. 2015. Systematic dissection of the sequence determinants of gene 3′ end mediated expression control. *PLoS Genet* **11:** e1005147. doi:10.1371/journal.pgen.1005147

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30:** 521–530. doi:10.1038/nbt.2205

Shen SQ, Myers CA, Hughes AEO, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel *cis*-regulatory analysis in the mammalian central nervous system. *Genome Res* **26:** 238–255. doi:10.1101/gr.193789.115

Shin J-E, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, Manglik A, Kruse AC, Marks DS. 2021. Protein design and variant prediction using autoregressive generative models. *Nat Commun* **12:** 2403. doi:10.1038/s41467-021-22732-w

Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, Nair S, Kundaje A. 2018. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. arXiv doi:10.48550/ARXIV.1811.00416

Shrikumar A, Greenside P, Kundaje A. 2019. Learning important features through propagating activation differences. arXiv doi:10.48550/arXiv.1704.02685

Siegel DA, Le Tonqueze O, Biton A, Zaitlen N, Erle DJ. 2022. Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *Adv Genet* **12:** jkab404. doi:10.1093/g3journal/jkab404

Smith C, Kitzman JO. 2023. Benchmarking splice variant prediction algorithms using massively parallel splicing assays. *Genome Biol* **24:** 294. doi:10.1186/s13059-023-03144-z

Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, Fairbrother WG. 2017. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet* **49:** 848–855. doi:10.1038/ng.3837

Sonenberg N, Hinnebusch AG. 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136:** 731–745. doi:10.1016/j.cell.2009.01.042

Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. 2017. Variant interpretation: functional assays to the rescue. *Am J Hum Genet* **101:** 315–325. doi:10.1016/j.ajhg.2017.07.014

Stark H, Jing B, Wang C, Corso G, Berger B, Barzilay R, Jaakkola T. 2024. Dirichlet flow matching with applications to DNA sequence design. arXiv doi:10.48550/ARXIV.2402.05841

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD®): 2003 update: HGMD 2003 update. *Hum Mutat* **21:** 577–581. doi:10.1002/humu.10212

Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA, Pourmand N, Sanford JR. 2013. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res* **23:** 1615–1623. doi:10.1101/gr.148585.112

Strayer EC, Krishna S, Lee H, Vejnar C, Beaudoin J-D, Giraldez AJ. 2023. NaP-TRAP, a novel massively parallel reporter assay to quantify translation control. bioRxiv doi:10.1101/2023.11.09.566434

Stroup EK, Ji Z. 2023. Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat Commun* **14:** 7378. doi:10.1038/s41467-023-43266-3

Tang Z, Koo PK. 2024. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. bioRxiv doi:10.1101/2024.02.29.582810

Taskiran II, Spanier KI, Dickmänken H, Kempynck N, Pančíková A, Can Ekşi E, Hulselmans G, Ismail JN, Theunis K, Vandepoel R, et al. 2024. Cell-type-directed design of synthetic enhancers. *Nature* **626:** 212–220. doi:10.1038/s41586-023-06936-2

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* **47:** D941–D947. doi:10.1093/nar/gky1015

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18:** 18–30. doi:10.1038/nrm.2016.116

Trauernicht M, Martinez-Ara M, Van Steensel B. 2020. Deciphering gene regulation using massively parallel reporter assays. *Trends Biochem Sci* **45:** 90–91. doi:10.1016/j.tibs.2019.10.006

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249:** 505–510. doi:10.1126/science.2200121

Uehara M, Zhao Y, Hajiramezanali E, Scalia G, Eraslan G, Lal A, Levine S, Biancalani T. 2024. Bridging model-based optimization and generative modeling via conservative fine-tuning of diffusion models. arXiv doi:10.48550/arXiv.2405.19673

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. 2021. Genome-wide association studies. *Nat Rev Methods Primers* **1:** 59. doi:10.1038/s43586-021-00056-9

Vainberg Slutskin I, Weingarten-Gabbay S, Nir R, Weinberger A, Segal E. 2018. Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nat Commun* **9:** 529. doi:10.1038/s41467-018-02980-z

Vainberg Slutskin I, Weinberger A, Segal E. 2019. Sequence determinants of polyadenylation-mediated regulation. *Genome Res* **29:** 1635–1647. doi:10.1101/gr.247312.118

Valeri JA, Collins KM, Ramesh P, Alcantar MA, Lepe BA, Lu TK, Camacho DM. 2020. Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat Commun* **11:** 5058. doi:10.1038/s41467-020-18676-2

Van Arensbergen J, FitzPatrick VD, De Haas M, Pagie L, Sluimer J, Bussemaker HJ, Van Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* **35:** 145–153. doi:10.1038/nbt.3754

Van Etten J, Schagat TL, Hrit J, Weidmann CA, Brumbaugh J, Coon JJ, Goldstrohm AC. 2012. Human pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. *J Biol Chem* **287:** 36370–36383. doi:10.1074/jbc.M112.373522

Verfaillie A, Svetlichnyy D, Imrichova H, Davie K, Fiers M, Atak ZK, Hulselmans G, Christiaens V, Aerts S. 2016. Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome Res* **26:** 882–895. doi:10.1101/gr.204149.116

Vilov S, Heinig M. 2024. Investigating the performance of foundation models on human 3′UTR sequences. bioRxiv doi:10.1101/2024.02.09.579631

Vlaming H, Mimoso CA, Field AR, Martin BJE, Adelman K. 2022. Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of RNA polymerase II elongation potential. *Nat Struct Mol Biol* **29:** 613–620. doi:10.1038/s41594-022-00785-9

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119:** 831–845. doi:10.1016/j.cell.2004.11.010

Wei Z, Hua K, Wei L, Ma S, Jiang R, Zhang X, Li Y, Wong WH, Wang X. 2023. neuronmotif: deciphering *cis*-regulatory codes by layer-wise demixing of deep neural networks. *Proc Natl Acad Sci* **120:** e2216698120. doi:10.1073/pnas.2216698120

Weile J, Roth FP. 2018. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Hum Genet* **137:** 665–678. doi:10.1007/s00439-018-1916-x

White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the *cis*-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci* **110:** 11952–11957. doi:10.1073/pnas.1307449110

Wissink EM, Fogarty EA, Grimson A. 2016. High-throughput discovery of post-transcriptional *cis*-regulatory elements. *BMC Genomics* **17:** 177. doi:10.1186/s12864-016-2479-7

Wong MS, Kinney JB, Krainer AR. 2018. Quantitative activity profile and context dependence of all human 5′ splice sites. *Mol Cell* **71:** 1012–1026.e3. doi:10.1016/j.molcel.2018.07.033

Wright CJ, Smith CWJ, Jiggins CD. 2022. Alternative splicing as a source of phenotypic diversity. *Nat Rev Genet* **23:** 697–710. doi:10.1038/s41576-022-00514-4

Wu X, Bartel DP. 2017. Widespread influence of 3′-end structures on mammalian mRNA processing and stability. *Cell* **169:** 905–917.e11. doi:10.1016/j.cell.2017.04.036

Wu Q, Bazzini AA. 2023. Translation and mRNA stability control. *Annu Rev Biochem* **92:** 227–245. doi:10.1146/annurev-biochem-052621-091808

Yartseva V, Takacs CM, Vejnar CE, Lee MT, Giraldez AJ. 2017. RESA identifies mRNA-regulatory sequences at high resolution. *Nat Methods* **14:** 201–207. doi:10.1038/nmeth.4121

Yin C, Hair SC, Byeon GW, Bromley P, Meuleman W, Seelig G. 2024. Iterative deep learning-design of human enhancers exploits condensed sequence grammar to achieve cell type-specificity. bioRxiv doi:10.1101/2024.06.14.599076

Yu Y, Maroney PA, Denker JA, Zhang XH-F, Dybkov O, Lührmann R, Jankowsky E, Chasin LA, Nilsen TW. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. *Cell* **135:** 1224–1236. doi:10.1016/j.cell.2008.10.046

Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ. 2014. Massively parallel functional annotation of 3′ untranslated regions. *Nat Biotechnol* **32:** 387–391. doi:10.1038/nbt.2851

Zhao J, Baltoumas FA, Konnaris MA, Mouratidis I, Liu Z, Sims J, Agarwal V, Pavlopoulos GA, Georgakopoulos-Soares I, Ahituv N. 2023a. MPRAbase: a massively parallel reporter assay database. bioRxiv doi:10.1101/2023.11.19.567742

Zhao S, Hong CKY, Myers CA, Granas DM, White MA, Corbo JC, Cohen BA. 2023b. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat Genet* **55:** 346–354. doi:10.1038/s41588-022-01278-7

Zheng W, Fong JHC, Wan YK, Chu AHY, Huang Y, Wong ASL, Ho JWK. 2023. Discovery of regulatory motifs in 5′ untranslated regions using interpretable multi-task learning models. *Cell Syst* **14:** 1103–1112.e6. doi:10.1016/j.cels.2023.10.011

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* **12:** 931–934. doi:10.1038/nmeth.3547

Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2019. A primer on deep learning in genomics. *Nat Genet* **51:** 12–18. doi:10.1038/s41588-018-0295-5

Zrimec J, Fu X, Muhammad AS, Skrekas C, Jauniskis V, Speicher NK, Börlin CS, Verendel V, Chehreghani MH, Dubhashi D, et al. 2022. Controlling gene expression with deep generative design of regulatory DNA. *Nat Commun* **13:** 5099. doi:10.1038/s41467-022-32818-8

Decoding biology with massively parallel reporter assays and machine learning

Alyssa La Fleur, Yongsheng Shi and Georg Seelig

Genes Dev. 2024, **38:** originally published online October 3, 2024
Access the most recent version at doi:10.1101/gad.351800.124