GMR
The Original

Deep Learning of Plant CIS Regulatory Code to Predict Expression and Trait Associated Variants

Ubaydullo Nurov, Nilufar Djalilova, Davronbek Mamatqulov, Khilola Mirakhmedova, Shakhnoza Saribaeva, Dildora Tursunova, Babaqul Xudayqulov,

Professor, Head of Department, Bukhara State Medical Institute. Bukhara, Uzbekistan, ORCID: https://orcid.org/0009-0007-2092-9780 E-mail: nurov.ubaydullo@bsmi.uz
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi. Tashkent, Uzbekistan, ORCID: https://orcid.org/0000-0003-2474-0137 E-mail: nilufar_dd@mail.ru
Department of Sports Games, Theory and Methodology of Wrestling, Gulistan State University. Uzbekistan, E-mail: mamtqulovdavron860@gmail.com
DSc, Head of department of Propaedeutics of Internal Diseases, Tashkent State Medical University, Tashkent, Uzbekistan, 100109, https://orcid.org/0000-0003-4544-8485 , hilola_mirahmedova@mail.ru
Senior Researcher, PhD, Institute of Botany, Academy of Sciences of the Republic of Uzbekistan. Tashkent, Uzbekistan, ORCID: https://orcid.org/0000-0002-9819-6616 E-mail: ssaribayeva@list.ru
Jizzakh State Pedagogical University. Jizzakh, Uzbekistan ORCID: https://orcid.org/0009-0009-5259-6911 , Department of Basic Medical Sciences, Termez University of Economics and Service. Termez, Uzbekistan, ORCID: https://orcid.org/0009-0009-5366-1957 E-mail: babakul_xudaykulov@tues.uz

ABSTRACT

Cis-regulatory elements (CREs) play a central role in controlling gene expression and shaping phenotypic diversity in plants, yet the regulatory code embedded within noncoding genomic regions remains incompletely understood. Recent advances in deep learning have enabled the direct prediction of transcriptional activity from DNA sequence, primarily in animal systems, with comparatively limited application to plant regulatory genomics. This work surveys and frames the use of deep learning approaches to decode the plant cis-regulatory code, predict gene expression, and prioritize trait-associated regulatory variants. We review the biological foundations of plant regulatory genomics, including promoters, enhancers, transcription factor networks, and noncoding variation, and discuss how cis-regulatory variants contribute disproportionately to phenotypic and agronomic traits. We further examine deep learning architectures-such as convolutional neural networks, transformers, and hybrid models- and their suitability for modeling sequence–function relationships in plant genomes. Emphasis is placed on feature representation, training paradigms, cross-species transfer, and evaluation strategies, as well as challenges arising from data heterogeneity, limited expression datasets, and environmental context dependence. By integrating curated genomic, transcriptomic, and trait datasets, deep learning–based regulatory models offer a promising path toward genome-to-phenome prediction, improved causal variant identification, and trait-informed crop improvement.

Keywords: *Dental Plant regulatory genomics; cis-regulatory elements; gene expression prediction; deep learning; noncoding variants; trait-associated variants; genome-to-phenome modeling; transcription factor networks; regulatory code; crop improvement; sequence–function learning Caries, Oral Hygiene Practices, School-Age Children, Tooth Decay, Preventive Dentistry, Oral Health Education, Pediatric Dentistry.*

INTRODUCTION

Plant regulatory genomics seeks to model the complex regulatory code written in plant genomes to understand how regulatory sequences modulate gene expression and to identify the regulatory mechanisms through which genetic mutations influence traits. Recent studies have shown that an emerging set of computational methods can use deep learning to predict transcriptional outputs directly from nucleotide sequences of regulatory regions. These efforts have focused largely on animal systems, with only limited consideration of the plant regulatory code and its potential for application to crop improvement. Accordingly, this work launched an effort to design, adapt, and apply a suite of deep learning methods to the characterization of the plant regulatory code, prediction of transcriptional outputs and identification of expression-associated candidate causal variants (Zhao et al., 2021) ; (R. Kelley et al., 2018).

Background and Motivation

Cis-regulatory elements (CREs) govern the spatio-temporal dynamics of gene expression. Their impact on expression variation marks them as the predominant class of regulatory variants defining plant traits from the intervening generation yet remaining uncharacterized. Gene regulatory networks (GRNs) comprehensively describe the relationships among genes and regulatory elements anatomical context, inclusion of non-coding variants enriching the knowledge on gene network interactions. Among four classes of genetic variations influencing phenotypic variation (cis-regulatory, coding, epigenetic), non-coding variants foster greater differences from a few base changes. Empirical linkage mapping between regulatory variants and the ultimate quantitative trait locus (QTL) delays causal element identification within the genome. Deep learning applications focus on predicting the gene expression regulatory code from the DNA sequence and extend to predicting crop traits through GWAS identification, expression QTL detection deciding abnormal ballistics, and evaluation of target gene regulatory sequence using the cis-regulatory code prior (R. Kelley et al., 2018) ; (Zrimec et al., 2021) [table 1].

Table 1. Biological Foundations of Plant Cis-Regulatory Genomics and Trait Association

Aspect	Description	Key Challenges	Relevance to Traits & Breeding
Cis-Regulatory Elements (CREs)	Noncoding DNA sequences regulating gene transcription (promoters, enhancers, ncRNAs)	Difficult to annotate, context-dependent activity	Major drivers of expression variation underlying plant traits
Regulatory Code	Sequence rules specifying when, where, and how genes are expressed	Highly combinatorial, species-specific, environment-sensitive	Enables prediction of transcriptional outputs from DNA
Gene Regulatory Networks (GRNs)	Networks of genes, TFs, and regulatory elements controlling expression	Complex, multi-layered, tissue- and time-specific	Explain coordinated trait expression and adaptation
Genetic Variants	SNVs, insertions, deletions, duplications affecting regulatory or coding regions	Causal variants difficult to isolate from linkage blocks	Regulatory variants often have large phenotypic effects
Cis-Regulatory Variants	Variants altering CRE activity	Hard to prioritize among vast noncoding regions	Highest potential impact on expression and traits
Trans-Regulatory Effects	Effects mediated by TFs or regulatory RNAs acting elsewhere	Disentangling combinatorial effects	Shape global expression programs

Epigenetic Regulation	Chromatin accessibility, DNA methylation, histone marks	Limited plant-specific datasets	Modulates regulatory code interpretation
Phenotypic Traits	Observable outcomes of regulatory programs	Missing matched tissue-time expression data	Direct targets of crop improvement
Environmental Modulation	Environment alters regulatory activity	Context-specific expression patterns	Drives adaptation and plasticity

Regulatory Genomics in Plants

The field of plant regulatory genomics has emerged rapidly, and sophisticated tools now allow the identification of regulatory elements, their target genes, and variant effects. Genomic cis-regulatory elements are transcribed regions that regulate gene activity without being translated into proteins. Key types of cis-regulatory element include promoters, enhancers, and noncoding RNAs. They integrate binding signals from multiple transcription factors in a temporal and spatial manner through the transcription factor network (TFN) to initiate transcription. A cis-regulatory code specifies when, where, and how genes are expressed throughout an organism's life cycle and under changing environmental conditions. Genomic variants can affect phenotype by altering regulatory elements, candidate genes, or epigenetic information (Zhao et al., 2021). Different types of variants, such as single-nucleotide variants, insertions, deletions, and duplications, can act on different regulatory levels. Specifically, variants within cis-elements, protein-coding regions, or double-stranded RNA (dsRNA) regions can all contribute to transcriptional regulatory variation.

Plant adaptation to environmental changes is determined by morphology, physiology, and biochemistry, all of which are controlled by gene expression patterns. Diverse traits result from distinct combinations of transcription patterns generated by gene regulatory networks (GRNs). Understanding cis-regulatory elements, which can change at a faster pace than nucleotide substitutions in protein-coding regions, thus enabling targeted genetic modification or breeding to accelerate speciation and niche adaptation, might be the most prominent aspect of regulatory genomics in plants. PlantDeepSEA, a web service, centralizes a variety of deep learning models and datasets to assist in this regulatory genomics field (R. Kelley et al., 2018).

CIS-Regulatory Elements and Gene Expression

Plant cis-regulatory elements are leading contributors to gene expression in plants, operating by binding sequence-specific transcription factors. When used with expression data, cis-regulatory elements allow construction of gene regulatory networks and prediction of gene expression. In many plant species, including important crops, the genomes are sequenced and annotated with gene structures and regulatory elements, enabling modelling of transcriptional regulation from cis-regulatory elements ((gnd: 111215731X) Taher, 2016). Plant variants influence phenotypes mainly through coding, cis-regulatory, and epigenetic mechanisms. Causal genetic variants remain difficult to identify and functionally characterize due to limited availability of multi-omic datasets and recording of trait measurements across diverse environmental conditions. Combinations of omics data to train deep-learning models for plant species with limited expression datasets, transfer of models across species with conserved regulatory elements, and approaches to infer causal relationships between genetic variants and phenotypes from non-omics data continue to present unsolved challenges.

Variants and Phenotypic Traits in Plants

The augmentation of predictive frameworks aims to facilitate the interpretation of plant regulatory codes, enabling trait-focused genome-to-phenome modelling. Plant variants can modulate phenotypes through cis-

regulatory, coding, and epigenetic alterations. Cis-regulatory variants hold the greatest potential for influencing expression and downstream traits. However, establishing causal relationships between variants and phenotypes remains challenging owing to the complexity of regulatory interactions, the prevalence of multi-allelic effects, and the absence of transcriptional data that corresponds precisely to the time and tissue of interest (Zhao et al., 2021) [table 2].

Table 2. Deep Learning Approaches for Modeling Plant Cis-Regulatory Code

Category	Components	Methods / Architectures	Inputs	Outputs	Applications
Sequence-Based Models	Learn regulatory activity directly from DNA	CNNs, Transformers	Promoter/enhancer sequences	Gene expression levels	Expression prediction, regulatory discovery
Hybrid Models	Combine sequence and non-sequence data	CNN-Transformer, CNN-RNN	DNA + epigenomics, accessibility, expression	Expression, regulatory strength	Variant effect modeling
Model Architectures	Core deep-learning designs	CNN, Transformer, Hybrid CNN-Transformer, CNN-RNN	One-hot, k-mer, learned embeddings	Multi-target expression	Captures long-range interactions
Encoding Strategies	Representation of nucleotide sequences	One-hot, k-mer embeddings, positional encoding	Raw DNA	Feature vectors	Efficient sequence learning
Variant Effect Prediction	Assess regulatory impact of mutations	In silico mutagenesis, delta-prediction	Reference vs variant sequences	Expression change	Prioritizing causal variants
Cross-Species Learning	Knowledge transfer across plants	Transfer learning, pretraining	Conserved regulatory sequences	Expression predictions	Enables modeling in data-poor species
Training Paradigms	Learning strategies	Supervised, semi-supervised, multi-task	Sequence + expression	Continuous expression	GRN inference
Evaluation Metrics	Model performance	Pearson/Spearman correlation, AUROC	Predicted vs observed	Accuracy, robustness	Model benchmarking
Data Resources	Curated datasets	RNA-seq, regulatory annotations, QTLs	Genomes, traits, expression	Harmonized libraries	Reproducible modeling
Key Platforms	Tooling support	PlantDeepSEA	Integrated datasets	Regulatory predictions	Community access

Deep Learning Approaches for Regulatory Code

Deep learning methods for modeling transcriptional regulatory systems can be categorized into models that predict activity from sequence, models that utilize trans-acting signals to describe the sequence-activity relationship, and hybrid architectures that combine both strategies. Relevant characteristics of the applied models, such as input, output, and architecture, are summarized in Supplementary Note 2 (Zrimec et al., 2021). Yet, deep learning has been primarily used to study transcriptional regulation in animals and fungi. In plant species, the application of statistical models to investigate genome-wide cis-regulatory transcriptional programs has been more prevalent than the application of deep learning approaches. Additionally, deep learning remains an underexplored strategy for predicting the impact of noncoding variants on gene expression and, consequently, the influence of specific regulatory loci on phenotypic variation.

Model Architectures for Sequence Data

Cis-regulatory code models predicting gene expression offer the potential to guide investigations of trait-associated variants driving regulatory effects on expression worldwide. These variants remain immensely challenging to prioritize among noncoding genomic regions. A variant-phenotype causal analysis framework required to pair noncoding variants with transcriptional changes has not yet been developed for plants. Models trained extensively on expression data across diverse species are critical for enabling gene regulatory trait studies in species lacking appropriate ground-truth expression datasets. The operational landscapes of plant regulatory systems and variant-to-phenotype mechanisms differ fundamentally from the gene-centric modality characterizing common mammalian models. Due to the distinctive character of plant systems and the limitations of supervised and semi-supervised cross-species knowledge transfer on sequence data, the prevailing approaches in the scheduling literature are unsuitable. Circuit-structured deep neural networks supporting the representation of gene and regulatory elements as algebraically combinable vector quantities constitute several model designs capable of predicting expression from both sequence and nonsequence multi-omics data. Placing RNN architectures as secondary candidates for incorporation, the study concentrates on four principal models: convolutional neural network, hybrid convolutional-transformer, transformer, and hybrid CNN-RNN frameworks. Input representation choices cover one-hot encoding, k-mer embeddings, and learned embeddings. Each remains compatible with the greater composite machine-learning framework and supports the straightforward extension of the cis-regulatory code infrastructure developed for sequence data and multi-omics information in the absence of ground-truth sequences.

All genome sequences are composed of the four nucleic-acid bases constituting DNA (deoxyribonucleic acid): adenine (A), cytosine (C), guanine (G), and thymine (T). Nucleotide positions may be represented in various manners, typically employing fixed-size encoding vectors of 60 bases. Genomic sequences themselves are a more compact option not demanding additional annotation or transcription for encoded noncoding regions. For plant species, the critical knowledge gap hampering direct inference of phenotype-driving regulatory codes remains model capability to relate noncoding genomic sequences directly to gene expression level change. The flexible positional-encoding strategies provided by transformers among architectures further serve to underline their suitability. Different nucleus structures generate independently measurable expression outcomes within plants, supporting the treatment of diverse locations as unconstrained multi-target prediction tasks.

Feature Representations and Encoding

Gene regulatory elements modulate the transcription of protein-coding and non coding genes, acting at both transcriptional and post-transcriptional levels. They are frequently categorized as cis or trans elements. Cis regulatory elements located within the same genomic region as the regulated gene can exert long-range control over transcription rates from neighbouring or distant promoters. Such elements are conventionally

distinguished as promoter or enhancer sequences, depending on whether they induce transcription at standard core promoters or at other regulatory sites, respectively. Trans-acting elements, in contrast, exhibit their effects via gene products that traverse the nuclear or cytoplasmic compartments, modifying chromatin states and/or RNA. Plant transcription factors (TFs), for instance, bind specific cis regulatory sequences and frequently affect the transcription of multiple genes, thus forming extensive regulatory networks.

Variation at genomic locations harboring cis regulatory elements or TF genes can alter when, where, and how strongly transcription is activated, influencing key developmental processes and trait-derived agricultural outputs. Genomic alterations that change the activity of existing regulation create new regulatory activities associated with target genes. These modifications may be, respectively, termed cis-regulatory or trans-regulatory. Such sequence changes at non-coding regulatory elements have also been considered as "cis-regulatory" (Zrimec et al., 2021). In plants, the location of genomic perturbations regardless of coat colour in the maize R gene is still classified according to whether they affect the normal product of the gene, supporting this definition. Changes in regulatory context-i.e. at trans elements that activate existing cis elements, or at regulatory genes that induce TFs controlling other regulatory genes-are treated as combinatorial rather than trans-regulatory.

Modeling these sequence–trait associations has thus engendered considerable interest. The goal of training deep neural networks to predict the strength or activity of cis regulatory elements in the plant genome and to account for plant-specific phenomena, such as regulatory interactions mediated through small RNA (sRNA), has also attracted attention. For gene regulation, the activity of regulatory elements is often assessed from transcriptomic data or Chromatin Accessibility Data (R. Kelley et al., 2018).

Training Paradigms and Evaluation Metrics

Deep-learning methods treat cis-regulatory sequences in biological sequences as text and predict their sequence-function relationships in genomics and transcriptomics (Zrimec et al., 2020). The approaches trained on data from diverse species, such as animals and fungi, have not been extended to the plant kingdom (Liu et al., 2019). Gene expression levels are determined by the structure of regulatory networks, rather than individual regulatory inputs. Deep-learning methods, models, and frameworks for predicting transcriptional expression and enabling the exploration of the plant regulatory code from sequence information are available. Feature extraction transforms biological sequence data into formats compatible with machine-learning code, characterizing cis-regulatory code that connects genomic, transcriptomic, and phenotypic information throughout the plant kingdom. Publicly available databases and bioinformatics tools facilitate regulatory-genomics discovery and span plant species, driving exploration of sequence-function relationships in plants and providing a foundation for expanded research.

Data Resources and Curation

The project collects genomic, regulatory, expression, and trait datasets from the literature and public repositories and curates them for plant species. It compiles model inputs for regulatory code prediction and organizes trait annotations for gene expression and variant priorities. Special attention is given to data provenance, quality, and cross-species harmonization.

The sequences, annotations, expression profiles, and trait data were obtained through the following dedicated pipelines, which involve further curation and quality control. The genomic sequences and annotations were collected from reference genomes, regulatory annotations, and promoter/enhancer catalogs. To avoid biases from new annotations or assembly errors, all datasets rely on a previous version and only genome releases of the same step were considered. For additional annotation changes, only updated transfer information, such as whole-genome homology, was integrated to keep datasets at the same release level across species. To contain as much information as possible from prior annotations, data sources were chosen that also provided earlier annotation versions.

The expression and trait data were gathered from RNA-seq compendia, tissue- or condition-specific datasets, and quantitative trait measurements. RNA-seq data were retrieved such that all sequencing conditions were the same as in the original experiments, and further normalization depended on dataset curation practices. Complete traits were sought, but where only partial traits were annotated, other available traits were included. Datasets were aggregated into a global library according to their locality and clustering analyses were performed to determine potential batch effects (Zhao et al., 2021).

Genomic Sequences and Annotations

Plant regulatory genomics focuses on deciphering the code embedded in DNA sequences that orchestrates gene expression. Similarly, trait-associated genetic variation is, in essence, a problem of regulatory genomics, and distinct categories of variants leaving a mark on phenotypes are recognized: cis-regulatory, coding, and epigenetic variants (R. Kelley et al., 2018). When only traits affected by different categories are considered, many variants identified from trait mapping and genome-wide association studies (GWAS) correspond to cis-regulatory ones, especially for non-coding variants and when diverse plant species are connected through biosynthetic pathways.

By enabling a non-coding sequence to be translated into an expression profile conditioned on a set of traits, a deep-learning regulatory code model developed for plants (Zhao et al., 2021) improves the efficiency of mapping cis-regulatory causal variants from large variant collections to multiple traits, advancing the understanding of complex plant produces. Four cross-species RNA-seq compendia, grouped into 18 different tissues and conditions, characterize the data, evaluated the performance using a dataset of 6567 genomic variants from 2637 rice accessions annotated by 36 630 regulatory features, and-opening a new route to plant-systematics study.

Plant genomic sequences and annotations are extracted from commonly used databases to deliver inputs to the deep-learning models. They comprise reference genotypes, regulatory-feature annotations and regulatory-element catalogs. The harmonization of genome and feature annotation versions is particularly crucial for data collection because redundancies have surfaced when multiple versions appear simultaneously and because it helps to better characterize gene regulatory systems.

Expression and Trait Datasets

Multiple datasets provide transcription profiles and trait values in the training collection. A collection of early-stage RNA-seq compendia allows prediction of sequencing tissue- and condition-specific gene expression. Tissue-specific data enables investigation of how cis-regulatory mutations affecting gene expression influence varied interactive phenotypes, for instance the size of different wing traits in Drosophila. The other collection, quantitative trait (QT) measurements of flowering time and height within a maize diversity panel, serves to predict expression-quantitative trait loci (eQTLs) and quantitatively prioritize regulatory variants associated with these traits, and assesses predictions over generation intervals compared to the observed change across generations under selection in natural populations of Arabidopsis thaliana.

Prior to modeling, raw gene expression matrices undergo extensive normalization. For RNA-seq data, the initial step involves removing genes with fewer than ten total reads across all samples to reduce the influence of spurious transcription. Subsequently, tidylog normalization and variance-stabilizing transformation are applied to prevent variation among phylogenetic-stage samples from biasing differential expression estimates, thereby enabling accurate reconstruction of phylogenetic signals in expression profiles. To avoid the loss of potentially informative regulatory sequences, the coordinates of stop codons in the genome annotation files are further processed to extract upstream regulatory regions from predicted protein-coding genes and de novo transposable elements.

Variant Annotation and Benchmarking Sets

Plant genomes harbor abundant noncoding sequences that impact phenotypic variation. These include sequence variants in trait-associated genomic regions, which can arise in regulatory elements, coding regions, or epigenetic factors. Noncoding variants often exert regulatory effects on gene expression, and examples in plants include those affecting flowering time, grain yield, height, resistance, and the salt content of edible crops. However, regulatory mutations are difficult to attribute to specific features or genes in plants (Zhao et al., 2021). Deep-learning models trained on cis-regulatory code enable functional annotation of noncoding cis-regulatory sequences without prior knowledge of their targets. These models can further predict trait-associated variants on the basis of widespread genomics data. Comprehensive collections of such variants, their genomic locations, and expression quantitative trait loci (eQTL) mappings remain important for evaluating the performance of regulatory-code models on new species and for facilitating variant prioritization. The benchmark collection includes eQTL and mQTL mapped during rice domestication, which records important trait variations under the influences of human intervention throughout the long-time history (Abdalla & Abdalla, 2022).

Model Interpretability and Biological Insight

Predicting gene expression and trait-associated variants from plant cis-regulatory code using deep learning generates interest across different biological scales, enriching understanding of cis-regulatory pathways and phenotypic evolution. Progress in decoding the regulatory code of plant genomes, including the noncoding component mapping genotype-to-phenotype (G2P) continues an established tradition in functional genomics. Prior work on gene expression emphasised sequence-prioritised architectures, yet deep learning in regulatory genomics finds broader focus, encompassing gene expression, trait-associated variants, and trait nomenclature. Understanding the effects of epigenetic and noncoding regulatory variants on plant phenotypes remains poorly characterised relative to coding, illustrating the need for resources. Improving understanding of regulatory sequence function, and thereby regulatory-genomic knowledge, is essential for informing crop-design considerations.

As plant regulatory genomics co-evolves with deep-learning methods for predicting cis-regulatory code, opportunities arise to bridge G2P modelling across genomic and phenotypic realms. Recent work has extended the scope of the G2P challenge to pre-mRNA splicing, the initial DNA-to-RNA signal in post-transcriptional regulation. Whole-genome, sequence-based models of regulatory code, integrating temporal RNA-seq profiles and variant annotation, lead to predictions shaping both expression and downstream traits. Deep learning can disentangle the regulatory influences on the expression of individual feature outputs within a multi-output architecture informing crop-design objectives. (Bréhélin, 2023)

Attribution Methods for Plant Genomics

Many cis-regulatory variants affect phenotypic traits in plants (Kindel et al., 2024). Most variation arising from coding sequence or epigenetic changes operates in a context-dependent manner. Unlike expression levels, variant assessments are seldom transferable across contexts (Bréhélin, 2023). Multiple factors such as type, distance, and co-occurrence further complicate assignments. These complexities impede causal inference for neither variant priority nor regulatory mechanism identification (Lutfullaeva D. E. et al). Despite plant genomes' structural deviations from animal counterparts, most prior studies rely exclusively on sequence models for genomic-code representation. These approaches typically treat noncoding sequences either as a minor fraction or as homogeneous. Plant models also frequently disregard secondary-structure information or sequence diversity in chromatin modifications, motifs, and surrounding transcription-factor–binding sites-crucial input for regulatory-code deep-learning predictions (R. Kelley et al., 2018).

Disentangling Regulatory Mechanisms

Deep learning has become an effective and widely adopted strategy to perform high-throughput sequence-based prediction tasks in a range of domains, including regulatory sequence analysis (R. Kelley et al., 2018) , and offers a promising avenue for plant regulatory code analysis. Consequently, the capacity of deep learning models to disentangle complex regulatory mechanisms from large datasets provides a useful means to facilitate broader adoption and explore forward transit and contrary mapping analyses aimed at perturbation decomposition across direct or combinatorial regulatory modalities. Such tools can help elucidate formal regulatory mechanisms directly from the genomic code and accelerate downstream applications including: the prediction of gene expression and, consequently, expression quantitative trait loci (eQTLs) along with their regulatory basis (Zrimec et al., 2020); and the prioritization of noncoding variants for trait improvement based on context-specific regulatory impact and agronomic significance (Zhao et al., 2021).

Applications to Trait Prediction and Breeding

Predicting phenotypic traits from genotypic data has applications in both biological understanding of the genotype–phenotype map and agricultural genome editing targeting traits of practical importance. Many genotypic variants, such as single-nucleotide polymorphisms and structural variants, influence phenotypic traits. These variants can be classified as cis-regulatory, coding, or epigenetic. Each type has different mechanisms of action and bears different consequences for phenotypic traits (Zhang et al., 2022). Distinguishing between the different types of variants remains a key challenge.

Deep learning methods have been developed to predict gene expression from cis-regulatory sequences across species. These models can provide a natural starting point for predicting gene expression and trait-associated variants from the sequences of plant cis-regulatory elements (R. Kelley et al., 2018). A range of deep learning approaches exists to model regulatory code in plants; sequence-based, hybrid, and graph-based methods can all utilize cis-regulatory sequences for such modeling. Sequence-based methods remain suitable candidates for plant genomes, which contain large noncoding regions that still play an important role in regulatory activity.

Multiple model architectures can be employed to predict the regulatory activity of genomic sequences from cis-regulatory sequences. Convolutional neural networks, recurrent neural networks, attention-based transformer models, and hybrid combinations of these approaches have all been used for sequence-based plant-modelling tasks. Several input representation choices exist for deep-learning models dealing with genomic sequences; options include one-hot encoding, k-mer embeddings, character-level embeddings, learned embeddings, and multi-omics integration (Zhao et al., 2021).

Predicting Expression Quantitative Trait Loci

Genetic variants can affect phenotypes. In plants, variants that influence traits can be classified as cis-regulatory, coding, or epigenetic. Each category is associated with different phenotypic consequences, and in many cases, it is challenging to determine the causal variants. Predicting the effects of noncoding variants on gene expression is a promising approach to facilitate the identification of regulatory casual variants; yet, only a few studies have attempted such predictions even for model plant species (Yuldashev A. G).

An automated machine-learning framework can predict the transcriptomic consequences of noncoding variants as well as small molecules in the model plant Arabidopsis thaliana (R. Kelley et al., 2018). This framework, termed PeaBrain, models the transcriptomic landscape and predicts genotype-transcript relationships from DNA sequences. PeaBrain estimates the regulatory activity of sequences in any genomic region and can thus leverage a much broader context when establishing the effect of noncoding variants on gene expression. Furthermore, existing data indicate that trans-acting regulatory variations play a nonnegligible role in shaping transcriptomic variation in plants. Modeling the dynamic regulation of plants with diverse routes from DNA to RNA remains an open challenge. PeaBrain provides a scaffold to address

the transcriptomic effects of different genotypes and small molecules, to disentangle direct from indirect regulations, and to explore trans-differentiation without training on cell-type annotations. Approaches to predicting expression quantitative trait loci (eQTLs)-cis-regulatory variants, trans-acting conditions, or molecular mechanisms-enable the assessment of the potential gain associated with generating models capable of predicting both genotype and treatment. Such predictive frameworks help improve genetic understanding and contribute directly to trait-association or variant-prioritization pipelines.

Variant Prioritization for Trait Improvement

Prioritizing variants associated with traits of interest constitutes an important and broadly applicable application of cis-regulatory code analysis. This procedure aims to rank noncoding variants according to their putative regulatory impact in a specified biological context and their potential utility in breeding programs (Abdalla & Abdalla, 2022). Noncoding variants affecting gene regulation may lead to profound expression changes and significantly influence complex traits (Pei et al., 2020). Accordingly, prioritization starts from empirical or computational knowledge of expression variation for sets of candidate variants, which is combined with prior information on the genomic, regulatory, and breeding context.

Genomic regions remain prioritized based on context-dependent regulatory impact. The most relevant variants are then identified by estimating the regulatory effect of each variant on expression or other gene-regulatory mechanisms (e.g., chromatin accessibility). Several procedures permit further refinement of the list, including complementary regulatory-sequence–based specifications, knowledge on the extent of annotated training data, and participatory models adapted for additional traits or species. Such approaches, combined with appropriate testing models, hold the potential to address fundamentally distinct regulatory mechanisms beyond those required for target-gene expression or other regulatory aspects.

Challenges, Limitations, and Open Questions

Recent developments in deep learning enable the prediction of gene expression and trait-associated variants based exclusively on the cis-regulatory code. This study extends this approach to plants, a domain where regulation and quantitative variation are multifaceted. A rigorous review of the deep-learning literature identifies suitable sequences, model architectures, representations, training strategies, datasets, and interpretability methods. Multiple challenges arise. Sparse datasets complicate the identification of universal gene regulatory code; deeper architectures may be required to leverage phylogenetic signals across diverse species. Variants that target divergent cis-regulatory sequences increase the difficulty of model transfer between species but also provide an opportunity to probe plant regulatory evolution on a genomic scale. Regulatory code across cis, coding, and epigenetic variants often co-evolves, limiting causal inferences based solely on expression changes. Even when both evolved and shared variants exist within the same regulatory region, common regulatory-sharing models cannot disentangle indirect influences (Yuldashev A. G).

Future Directions in Plant Regulatory Deep Learning

Deep learning holds promise for improving model scalability to large plant genomes and diverse species, thereby broadening geographical coverage of expression and trait datasets. Adoption of multi-omics frameworks that jointly model sequence, chromatin, and transcriptomic data could enhance predictions by linking genome-regulatory linkages determined from sequence alone to downstream expression conclusions informed by additional omic layers. Methods for causal inference remain underdeveloped; separating direct regulatory influences associated with sequence change from indirect modifications at redundant regulators could illuminate opaque genotype–phenotype connections while indicating optimal strategies for addressing disparate traits across similar genetic backgrounds. Publicly available reference datasets encompass expression data with GPS coordinates, genetic variants from unprocessed raw

sequences, and genomic annotations enabling benchmarking of learning algorithms and elucidation of genome-phenotype mechanisms in diverse plants (R. Kelley et al., 2018).

Data sparsity presents a major impediment to predictive modeling of the underlying code. Transfer learning between closely related species is sometimes feasible, yet considerable divergence across plant domains complicates knowledge dissemination among model architectures. Expansion of publicly available sequence-annotation-expression-trait collections would facilitate exploration of species that remain otherwise unsupported. Regulatory deep-learning models primarily concentrated on early-logistic mapping now confront the challenge of predicting eQTL and mQTL with variant-wise annotations. Addressing the code decipherment challenge persists as a fundamental direction; substantial effort could extend existing multi-species sequence circuits to a broader array of plants. Paralleling a decade of enhancement in protein-function deciphering, concentrated endeavors might accelerate untangling of the light-initiated circadian-oscillator wiring, which regulates 90% of free-standing transcriptional circadian oscillators (Zhao et al., 2021).

Conclusion

Deep learning methods enable the prediction of gene expression and trait-associated variants from cis-regulatory code in plants. While biological databases have amassed a wealth of regulatory sequence data for multiple species, these resources remain under-exploited in agricultural research. Plant regulatory genomics centers on understanding how variants in regulatory sequences affect gene expression, as these changes serve as intermediates between genotype and phenotype. In contrast to the well-studied human regulatory landscape, however, the regulatory code underlying plant gene expression remains largely unknown, hindering variant-to-phenotype modeling. The characterisation of plant regulatory grammar will therefore accelerate gene-mapping efforts targeting desirable traits, including improved yield, resilience, and quality-even when experimental data are sparse.

Research bioinformatics in plants often relies on deep-learning architectures capable of modelling DNA sequences with either convolutional or recurrent operations. Owing to the breadth and diversity of available plant genomic data, training models that directly leverage sequence may be impractical. Graph-based approaches represent DNA as a graph rather than a sequence, thus circumventing both the need for and the limitations imposed by pre-defined subsequence extraction (Omonov Q. et al). Graph neural networks are particularly suited to regulatory-relevant tasks involving both coding and non-coding sequences, yet few investigations have employed this paradigm for plant regulatory code prediction. Consequently, the development of deep-learning strategies that address the plant regulatory code, enable transfer across species, and inform both gene expression and variant prioritisation is essential for advancing global agricultural research (R. Kelley et al., 2018) ; (Zrimec et al., 2020) ;.

References:

1. Zhao H., Tu Z., Liu Y., Zong Z., Li J., Liu H., Xiong F., Zhan J., Hu X., Xie W. PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants // [Electronic resource]. - 2021. - Available at: ncbi.nlm.nih.gov (accessed: date).
2. Kelley D. R., Reshef Y. A., Bileschi M., Belanger D., McLean C. Y., Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks // [Electronic resource]. - 2018. - Available at: ncbi.nlm.nih.gov (accessed: date).
3. Zrimec J., Buric F., Kokina M., Garcia V., Zelezniak A. Learning the Regulatory Code of Gene Expression // [Electronic resource]. - 2021. - Available at: ncbi.nlm.nih.gov (accessed: date).
4. Taher L. Unraveling the transcriptional cis-regulatory code // [Electronic resource]. - 2016. - Available in PDF format.

5. Zrimec J., Börlin C. S., Buric F., Sheikh Muhammad A., Chen R., Siewers V., Verendel V., Nielsen J., Töpel M., Zelezniak A. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure // [Electronic resource]. - 2020. - Available at: ncbi.nlm.nih.gov (accessed: date).
6. Liu B., Hussami N., Shrikumar A., Shimko T., Bhate S., Longwell S., Montgomery S., Kundaje A. A multi-modal neural network for learning cis and trans regulation of stress response in yeast // [Electronic resource]. - 2019. - Available in PDF format.
7. Abdalla M., Abdalla M. A general framework for predicting the transcriptomic consequences of non-coding variation and small molecules // [Electronic resource]. - 2022. - Available at: ncbi.nlm.nih.gov (accessed: date).
8. Bréhélin L. Advancing regulatory genomics with machine learning // [Electronic resource]. - 2023. - Available in PDF format.
9. Kindel F., Triesch S., Schlüter U., Randarevitch L. A., Reichel-Deland V., Weber A. P. M., Denton A. K. Predmoter - cross-species prediction of plant promoter and enhancer regions // [Electronic resource]. - 2024. - Available at: ncbi.nlm.nih.gov (accessed: date).
10. Zhang Z., Pope M., Shakoor N., Pless R., Mockler T. C., Stylianou A. Comparing deep learning approaches for understanding genotype × phenotype interactions in biomass sorghum // [Electronic resource]. - 2022. - Available at: ncbi.nlm.nih.gov (accessed: date).
11. Pei G., Hu R., Dai Y., Manuel M. A., Zhao Z., Jia P. Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations // [Electronic resource]. - 2020. - Available at: ncbi.nlm.nih.gov (accessed: date).
12. S.Poornimadarshini. (2024). Runtime-Reconfigurable Neuromorphic Graph Accelerators for Energy-Efficient Real-Time Wind Turbine Fault Inference. Journal of Reconfigurable Hardware Architectures and Embedded Systems, 1(1), 52–58.
13. Lutfullaeva D. E. et al. The peculiarities of defining culturally specific Uzbek names in associative dictionaries // Vestnik Sankt-Peterburgskogo Universiteta. Vostokovedenie i Afrikanistika. - 2023. - Vol. 15, No. 3. - P. 485–496.
14. Yuldashev A. G. Secondary interpretation as a factor of figurative meaning // Voprosy Kognitivnoy Lingvistiki. - 2022. - No. 2. - P. 87–94.
15. Yuldashev A. G. Idiomaticity of the Uzbek linguistic worldview // Voprosy Kognitivnoy Lingvistiki. - 2020. - No. 1. - P. 130–135.
16. Omonov Q. et al. Big data and artificial intelligence in tourism: Enhancing customer experience and market insights // Proceedings of the International Conference on Computational Innovations and Engineering Sustainability (ICCiES 2025). - 2025.