Julian McKinley

BSC4943r

Dr. Karen McGinnis

Summer 2024

# Binding Interactions Between Long non-coding RNA and RNA Binding Proteins Influence the Regulation of Genes Through Chromatin Remodeling in *Zea mays*

## Abstract

Gene regulation is a process that controls the expression of genetic information and is essential for all forms of life. In plant genetics, gene regulation is complex, multifaceted, and is often influenced by interactions between long non-coding RNAs (lncRNAs) and RNA-binding proteins (RBPs). This research project investigates these interactions in *Zea mays* (maize), a plant species that not only demonstrates these genomic characteristics but has been well-studied in the field of genetics for about a century. The main objective in this research was to predict the binding interactions between maize lncRNAs and RBPs that contribute to gene regulation, while also considering the possible influence of chromatin structure alterations and their corresponding chromatin states. Utilizing bioinformatic tools like OrthoDB, ChromHMM, and RPISeq and databases like MaizeGDB, PCSD, RBPsuite, the influence of chromatin structure and the binding affinities of these interactions were computationally analyzed. This project implemented a comparative transcriptomic approach because it used human CLIP-seq data as a model since extensive CLIP-seq data is not as available for maize. Additionally, maize eQTL data was used to identify potential regulatory targets of lncRNAs and RBPs in maize that were chosen testing.

Using RPISeq, a web-based bioinformatics tool, I predicted whether binding occurs between maize versions of human RBPs (FKBP4, TAF15, and WDR43) and lncRNAs (GRMZM2G476477 and GRMZM2G018006), along with other lncRNAs that regulate the target genes with similar chromatin state arrangements as those regulated by the two lncRNAs listed above. My findings are that these interactions can be difficult to categorize based on chromatin state arrangement alone, but there does appear to be some sort of chromatin response. There is a distinct difference between the chromatin states present in lncRNAs and the chromatin states present in the genes regulated by lncRNAs. This distinction works to either suppress or express genes in maize, depending on which states are involved during RBP binding.

## Background

### Genetic Terms and Research Approaches

The fundamental element of inheritance transmitted from one generation to the next is known as the gene. Genes hold instructions that synthesize proteins and perform cellular functions (National Cancer Institute). The entire set of genetic information or DNA in an organism is referred to as the genome and the study of the structure and functions of genes within its genome is known as genomics. Genomics is concerned with understanding the genome and mapping genes to their corresponding DNA (National Human Genome Research Institute). Transcriptomics, a subgroup of genomics, focuses on the transcriptome, which is the complete set of RNA molecules (transcripts) and their corresponding genes (Anthony & Hondermarck, 2023). Transcriptomic approaches provide insight on the expression and repression of genetic information by concentrating on RNA molecules and their roles in the transcriptome. By studying the transcriptome, we can better understand gene regulation and the genomic structures of species of organisms and life on our planet as a whole (Wang et al., 2009).

## The Central Dogma

In the past, biologists have approached genomic and transcriptomic research with a theory that is now considered to be oversimplified. This theory, known as the central dogma of molecular biology, describes the foundation for how genetic information flows within a biological system. Proposed by Francis Crick in 1958, it states that genetic information only flows from DNA to RNA to protein. This unidirectional flow consists of the processes of transcription, where DNA is transcribed into RNA, and translation, where RNA is translated into proteins (Crick, 1958). Although the central dogma remains in its truth and has been widely accepted as a fundamental principle in molecular biology, more recent research has not only showed that there are many additional elements that impact how genes are expressed in proteins, but there are also other biological processes that accompany this theory. For instance, the discovery of retroviruses revealed that with the use of the enzyme, reverse transcriptase, these systems transcribe RNA into DNA, challenging the previously held notion that all genetic material proceeds unidirectional (Coffin, 1997). Additionally, it is now understood that most RNAs do not code for proteins, but instead play a crucial role in gene regulation. This discovery reveals that the processes of gene expression and protein synthesis are much more complex than previously thought. Though this does not change our grasp of gene expression and protein synthesis, it does add to our understanding of the central dogma itself and how these processes are regulated.

## RNA and Protein Synthesis

RNA, or ribonucleic acid, is a polymer made up of nucleotides, whose primary function, in accordance with the Central Dogma, is to make proteins. Each nucleotide consists of three components: a sugar molecule (ribose), a phosphate group, and a nitrogenous base (adenine, guanine, cytosine, or uracil). These nucleotides are linked together through phosphodiester bonds (Wang, 2023). Protein synthesis involves three primary RNA types: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). Messenger RNA (mRNA) is produced from DNA and carries the genetic instructions for protein synthesis known as transcripts. In prokaryotes, mRNA is immediately available for protein synthesis, while in eukaryotes, it undergoes processing to become mature mRNA and gets transported from the nucleus to the cytosol so that it can eventually be translated into a protein. Transfer RNAs (tRNAs) provide amino acids for translation to take place. By transporting amino acids to the ribosome, tRNAs assist in the translation of mRNAs into proteins. tRNA's anticodon arm pairs with mRNA codons to ensure the correct amino acid is brought to the ribosome. Ribosomal RNAs (rRNAs) are responsible for the structure and catalytic activity of ribosomes, the cell-site where protein synthesis takes place. Ribosomes have specific sites (E, P, A) for binding mRNAs and tRNAs, ultimately linking amino acids to form polypeptides and hold the growing peptides together (Wang, 2023).

## Regulatory Roles of Noncoding RNA

In addition to the role of RNA described by the central dogma, latest discoveries in research have unveiled some of the numerous regulatory roles of a type of RNA called noncoding RNA (ncRNA). Making up 98-99% of the human transcriptome, ncRNAs were initially thought to be background noise, having no role in the process of gene expression and regulation. However, it has been well-established in recent years, that these RNA molecules may not encode for proteins, but they perform many functions in the cell that are crucial for controlling gene expression (Kaikkonen et al., 2011). When analyzing the three primary RNA molecules for protein synthesis, we can categorize both tRNAs and rRNAs as ncRNAs and mRNA as the coding RNA molecule, although, some ncRNAs have regulatory functions that are completely separate from translation. Over 40 different types of ncRNAs have been discovered, the vast majority being tRNAs and rRNAs (Pranati & Sarat, 2021). However, another type of ncRNA that is frequent in the transcriptome, known as long non-coding RNA (lncRNA), was considered for this research project because of its relatively recent discovery and its emerging understanding of its role in modulating gene expression.

### A Type of ncRNA: Long non-coding RNA

Exceeding 200 nucleotides in length, lncRNAs are found in both the nucleus and the cytoplasm (Kaikkonen et al., 2011). LncRNAs are noted for their roles in regulating gene expression during transcription, after transcription has occurred, and during translation. They do this primarily by forming interactions with DNA, mRNA, and proteins (Zhang et al., 2019). Through these interactions, lncRNAs can control transcriptional processes by influencing chromatin structure or alterations in histone proteins. They can exert post-transcriptional regulatory roles by regulating splicing, facilitating mRNA degradation, or preventing translation from taking place. In translation, they can serve as protein activity modulators, scaffolds, or as decoy receptors (Bhattacharyya et al., 2021).

### Challenges with ncRNAs and lncRNAs

NcRNAs overwhelm the human transcriptome, whereas only 1-2% of RNA sequences encode for proteins (Kaikkonen et al., 2011). Despite the vast occurrence of this phenomenon, we know relatively little about ncRNAs, specifically lncRNAs. Researchers are continuing to classify all the functions of lncRNAs and to understand the extent to which these molecules influence gene expression. Advancing our understanding of lncRNAs and their regulatory functions is a novel task that can be considered a challenge to researchers, due to various factors related to their complex nature. The first has to do with gene conservation. The rules of conservation for protein-coding RNAs are well established but they may not apply to lncRNAs. Unlike protein-coding genes, the functions of lncRNAs cannot be interpreted by their primary sequence alone. This does not mean that lncRNAs do not have a conserved sequence or structure, but it might imply that they conserve their structure differently and more complex than protein-coding RNA sequences. Another challenge is that lncRNAs may have unique roles in various cellular processes, making it hard to pinpoint specific functions associated with distinct lncRNAs. Furthermore, it is difficult to assign structural and functional relationships when analyzing a big group of molecules with multiple, diverse functions all lumped into one group. Additionally, some lncRNAs might not even have any regulatory roles, creating background noise and making their classification even more of a challenge. Another factor related to their complex nature is that lncRNAs are expressed at low levels, making them harder to detect and quantify accurately. The subcellular localization of lncRNAs can also vary, with some found in the nucleus and others in the cytoplasm. This can influence or reflect their functions and mechanisms of action. Due to the large number of lncRNAs and their diverse functions, there can even be functional redundancy, where multiple lncRNAs may have overlapping or redundant roles, hindering the determination of the specific functions of individual lncRNAs (Mercer et al., 2009).

These factors related to their complexity not only emphasize the difficulty in doing research on these molecules, but also stress the importance of both advancing our understanding of lncRNAs and identifying their potentially extensive undiscovered regulatory functions. Their overwhelming presence in the transcriptome, tied in with our incomplete comprehension, has motivated me and many other researchers to pursue these areas further. To achieve this, researchers have recognized the role of RNA-binding proteins (RBPs) in providing valuable insights into specific regulatory functions of lncRNAs.

**RNA-binding Proteins (RBPs)**

RNA-binding proteins (RBPs) are a set of proteins that bind to several classes of RNA molecules, including lncRNAs, and are essential for RNA metabolism and RNA function. Through structural motifs and domains, they interact with RNA molecules by forming complexes that influence gene expression and cellular function. There are several ways that lncRNA and RBP binding interactions can impact regulation. RBPs can bind with lncRNA molecules and direct them to specific locus points, where the lncRNAs are able to recruit chromatin-modifying complexes that authorize the binding of additional regulators. These complexes can be recruited to the sites of lncRNA transcription or away from the sites, near genes that are targets for regulation. LncRNAs can act as decoys for RBPs, inhibiting availability for their typical genomic targets and potentially influencing chromatin structure, causing indirect regulation. RBP complexes can also be formed with lncRNAs that influence the promoter regions of genes targeted for regulation (Batista & Chang, 2013). Three examples of RBPs that bind to lncRNAs in the human transcriptome are FKBP4, TAF15, and WDR43. FKBP4 belongs to the immunophilin family and is involved in fundamental cellular processes such as protein folding and trafficking (NCBI). TAF15 is a TATA-box binding protein that is part of the TET family of RBPs. It’s involved in RNA polymerase II gene transcription, functioning as part of transcription initiation factor complexes (NCBI). WDR43 facilitates RNA polymerase II complex binding and regulates transcription elongation (NCBI). These three human RBPs are a few examples of how we can tie lncRNAs to a specific regulatory function by referencing the functions of the proteins involved in the binding interactions. By understanding the interactions of specific RBPs and specific lncRNAs, it can help give us insight into the types of regulation being applied. In other words, knowing what kinds of RBPs are involved and how they perform their regulatory functions, may allow us to predict the regulatory functions of certain lncRNAs that form interactions with particular RBPs. To gain further insight on this phenomenon, RBPs that were predicted to regulate the same loci as certain lncRNAs were used for this research project. Moreover, the structure of chromatin and its impact on gene regulation, became a specific regulatory function that was of interest to see if it was influenced by certain lncRNA and RBP binding interactions.

## Human Genetics and CLIP-seq Data

In human genetics, there has a been a massive effort to understand and assemble our genome. The Human Genome Project began in October 1990 and was completed in April of 2003. Regarded as one of the greatest scientific accomplishments in history, the Human Genome Project generated the first human genome, providing our entire assembly of genetic information and ultimately paving the road for advancements in understanding mammalian genetics (National Human Genome Research Institute). In addition to the Human Genome Project, there has been an immense bioinformatic accumulation of genomic and transcriptomic sequencing data oriented towards human and mammalian genetics over the past few years. One type of transcriptomic data known as CLIP-seq data, which is RNA sequencing data that contain the RNA sites and their corresponding binding proteins, proved to be crucial for the research on this type of regulation and the computational analyses demonstrated throughout this project (Hafner et al., 2021).

CLIP-seq is just one of many kinds of human transcriptomic sequencing data that became recently available for the public. Cross-linking and immunoprecipitation (CLIP) approaches focus on binding proteins, identifying RNA sites bound by RBPs. These techniques involve the UV irradiation of cells, causing proteins near the irradiated bases to form covalent bonds with RNA. This allows for the stringent purification of RNA-protein complexes and the identification of the interactions of specific proteins across the transcriptome. Additionally, CLIP uses RNase treatment of cross-linked protein complexes to isolate RNA fragments bound by the RBP. Sequencing these fragments identifies RBP binding sites, allowing for the prediction of RBP function (Hafner et al., 2021).

**Plant Genetics, *Zea mays,* and Comparative Studies**

Although human genomics and transcriptomics has been a focal point in the field of genetics, it's important to recognize that pursuing genetic research on other species, like certain plants species, can provide valuable insights into the mechanisms involved in regulating gene expression. In addition, plant genetic research is important not only because plants play vital roles for humans and various other organisms by serving as food and medication, but they can be easily manipulated in lab environments, are usually more accessible for experiments, and often involve fewer ethical concerns when doing genetic research (Ong et al., 2016). *Z. mays* serve as a great model for exploring the control of gene expression not only for all these reasons, but also because of its genome rich in transposable elements and the extensive genetic research done on maize in the past (J. Huang a et al., 2016). *Z. mays* has been used tremendously to study generational inheritance, including the most fundamental principles in the field of genetics: the laws of segregation, independent assortment, and dominance. Gregor Mendel, renowned as the “father of genetics”, used corn as a model to study these laws and establish the concepts recognized by geneticists today (Eckardt et al., 2022). Overall, maize has been used for all sorts of genetic research for over a century, especially in Mendelian genetics, meaning it’s a plant species that’s been thoroughly studied, having lots of data and information (Southern Biological).

Even though maize is a plant species that’s been well-studied and is recognized as an excellent model for researching gene regulation, there is still significantly more published genomic and transcriptomic data for humans. Therefore, extensive CLIP-seq data is not as available for *Zea mays* or plant species in general. This could be due to a lack of funding in the field of plant genetics as well as the complex structure of plant genomes. Genetic information in plants is organized differently than in mammalian genomes, often having high ploidy levels, repetitive sequences, and frequent genome rearrangements, which can pose challenges in plant genomic and transcriptomic research. However, through comparative studies, this issue can be resolved. Comparative studies essentially use data from other species or organisms as a model to compare to the species or organism of interest. These types of studies can provide valuable insights into gene conservation, regulation, evolution, and functional diversity across different species that lack certain kinds of data (Proost et al., 2015). Comparative genomics and transcriptomics involve the comparison of species with well-characterized genomes and transcriptomes, such as humans or mammals, to those with less extensive bioinformatic data, like *Z. mays* and other plant species (National Human Genome Research Institute). For this research project, a comparative transcriptomic approach was employed by leveraging human CLIP-seq data to investigate the binding interactions between lncRNAs and RBPs in maize. The study also aimed to determine whether changes in chromatin structure occurs, if these changes are due to the formation RBP-complexes, and if this could be considered a specific function of lncRNAs.

## Bioinformatic Databases and Tools

### MaizeGDB

The Maize Genetics and Genomics Database (MaizeGDB) was utilized in this project due to its vast repository of maize-related data and its collection of tools useful for data analyses. As the central hub for published data in maize, MaizeGDB offers a wide variety of resources, including sequences, chromosomal mapping, gene models, and several other options (Lawrence et al., 2004). MaizeGDB was essential for obtaining all sequences of the RBPs and lncRNAs that were tested for predicting binding interactions. In general, the platform played a crucial role in the bioinformatic analyses carried out for this research project.

### RBPsuite

RBPsuite, a computational framework that provides tools and libraries containing proteins that bind to RNA, was used to obtain the human CLIP-seq data so that a transcriptomic comparative study could be implemented. RBPsuite contains benchmark human CLIP-seq data of 154 RBPs for linear RNAs, which is available for public use (RBPsuite). RBPsuite uses iDeepS, a deep learning model that predicts RBP binding sites on linear RNAs, to essentially generate the CLIP-seq data. The iDeepS model is trained on downloaded peaks for 154 RBPs of K526 and HepG2, corresponding to the human genome hg38 version. These narrow peaks were generated using the eCLIP-seq Processing Pipeline v2.0 of ENCODE for humans. This results in the generation of benchmark, or published, CLIP-seq data for 154 human RBPs (RBPsuite).

### eQTL Data

Aside from using CLIP-seq data from the human transcriptome, maize transcriptomic data from an expression quantitative trait loci (eQTL) analysis was also used for research and acquired from a published database (Fu et al., 2013). This type of analysis associates genetic variation that appears to be linked to variable expression by looking at large, genetically diverse populations of species. From this, regulated genes are predicted and associated with genomic regions that appear to encode regulators responsible for regulating the target genes. The regulator regions could be DNA sequences, acting as promoters or enhancers, or proteins and RNA sequences, acting as regulatory factors. Moreover, eQTL maize data was necessary to from the subgroups of the RBPs and lncRNAs chosen for testing the binding affinities.

### OrthoDB

OrthoDB is one of the computational tools that facilitated a comparative transcriptomic approach for this project. The database serves as an archive for identifying protein-coding genes shared among diverse species and organisms. These shared genes are referred to as orthologs because they diverge into separate genes and species and originate from a common ancestor. Orthologs are a result of evolution, demonstrating functions and structures that are conserved across genomes and transcriptomes. OrthoDB works by collecting translation data from extensive, varying species databases like GenBank and RefSeq. Using alignment tools such as BLAST, OrthoDB classifies these genes into orthologous groups, which are classes of genes that evolved from a common ancestral gene by means of speciation, in which a new population of organisms is created (National Geographic). These groups allow for the identification of similarities and evolutionary relationships among genes across species. It then arranges these groups in a phylogenic tree that illustrate the evolutionary changes between species, along with supplying detailed gene annotations (Kuznetsov et al., 2022). Although OrthoDB primarily focuses on orthologous relationships, the platform can also provide information about paralogous groups, which are classes of genes that diverge from one another because of gene duplication. Orthologs and paralogs are two distinct kinds of gene separation caused by evolution, both of which fall under the broader category of homology. Homologs are genes related through evolution, and can be orthologs, paralogs, or simply just homologous genes if its reasons for similarity are unspecified. (National Library of Medicine). This tool is an excellent resource for making comparative analyses with genomes and transcriptomes of other species. For this project, OrthoDB was useful in identifying maize RBPs' homologs by comparing them to human RBPs identified through CLIP-seq data.

### ChromHMM and PCSD

Although much is yet to be revealed about lncRNAs, one of the roles that biologists recognize in lncRNAs, is its impact on gene regulation, particularly their influence on chromatin structure. By using a bioinformatic tool known as ChromHMM and Plant Chromatin State Database (PCSD), we can analyze the varying structures of chromatin in maize and determine if it is regulatory function influenced by lncRNA and RBP binding. ChromHMM is a tool in computational biology used to infer chromatin structures in plant genomes, while PCSD is a comprehensive database that provides and collects information about chromatin structure in plant genomes (Liu et al., 2018). In terms of gene regulation, we know that chromatin can change its structure to either repress or express certain genes. ChromHMM recognizes this and classifies each varying chromatin structure into what is known as chromatin states, by integrating histone modification data and other epigenomic data. These chromatin states are arranged from compact states to states accessible by transcription, enabling replication, repair, and recombination (Tabassum & Parvez, 2021). ChromHMM works by using chromatin immunoprecipitation sequencing (ChIP-seq), which indicate the varying histone modifications and chromatin marks across a plant genome. It assigns sections of the genome to certain chromatin states based on the presence of each histone modification and mark, as well as their distinct patterns of arrangement. Ultimately, ChromHMM characterizes chromatin states, each having different functions, then stores this information in PCSD, effectively complementing each other to further enhance the study of chromatin in the field of plant genetics (Ernst & Kellis, 2012).

When using ChromHMM to research chromatin structure in *Z. mays*, it finds that they display 26 different chromatin states, each corresponding to varying levels of gene expression. In the maize transcriptome, it’s possible that RBPs can bind to specific lncRNAs and regulate a certain gene by remodeling chromatin near the locus of that gene into a specific arrangement of those 26 states. LncRNAs can recruit RBPs that form chromatin-modifying protein complexes directly changing states, interact with RBPs that prevent the formation of protein complexes, and form chromatin loops by binding to proteins (Tabassum & Parvez, 2021). By categorizing chromatin states in maize and mapping the arrangement of states to the genes regulated by lncRNAs and RBPs, we might be able to better understand the regulation roles of lncRNAs and how they may influence chromatin structure.

### RPISeq

RPISeq is the web-based tool that was used to predict the binding affinities for the lncRNAs and RBPs in maize. Using only sequence information, RPISeq was designed to predict RNA-protein interactions. The RNA sequence is represented as a normalized vector of its ribonucleotide 4-mer composition, and the protein sequence is represented as a normalized vector of its 3-mer composition, using a 7-letter reduced alphabet representation. Additionally, RPISeq uses Support Vector Machines (SVM) and Random Forest (RF) machine learning models, which are trained on RPIntDB, to make these predictions. RPIntDB is a database with known RNA-Protein interactions, containing a total of 30,056 of these interactions. These interactions are known because of the relatively recent emergence of high throughput sequencing in the field of computational biology, which made these identifications. However, these are time consuming and expensive, which is why RPISeq was developed and makes quick predictions using available benchmark data of known RBP interactions (Muppirala et al., 2011). SVM models are tasked to make classifications, which use a decision boundary to split the data into different classes based on varying characteristics or outliers (1.4. Support Vector Machines). Random Forest models create multiple subsets of the data it's originally trained on, trains the data on those subsets to consider the features of each subset, and continually splits the data into subsets. The outputs of the multiple subsets generated, are then averaged to make its predictions (What is Random Forest?, 2021). The scores generated in RPISeq from SVM and RF classifiers, are represent by probability values ranging from 0 to 1. Values above .5 represent a positive binding score but have varying levels of prediction confidence. Scores closer to 1 indicate high confidence binding, scores between .5 and .7 represent moderate binding confidence, and scores below .5 are categorized as low confidence predictions. RPISeq was utilized for predicting binding affinities and served as a simple platform to generate rapid results.

## Goal

The goal of this project was to predict whether certain lncRNAs in *Zea mays* can regulate the expression levels of certain genes by binding to proteins and inducing certain chromatin state arrangements. The aim was to bridge the knowledge gap in gene expression control in maize through lncRNA regulation, thereby enhancing our overall understanding of noncoding RNAs and their role in the operation of plant genomes and transcriptomes. Additionally, this project sought to draw parallels between plant and mammalian genetics. Utilizing CLIP-seq data from humans, a comparative transcriptomic study was conducted to investigate the regulatory influence of lncRNA and RBP binding interactions in maize.

## Methods

### Comparing Human ortholog to *Zea mays*

Due to the lack of extensive maize CLIP-seq data, CLIP-seq data for 154 human RBPs was obtained from the RBPsuite online database (RBPsuite). The corresponding maize RBPs were identified based on similarity to the humans RBPs. Gene IDs and accession numbers for all 154 of the human RBPs were identified and retrieved using NCBI. Using this information, OrthoDB was implemented to map the homologs and locus IDs in *Z. mays.* The homologous RBPs identified were then used to find their associated gene IDs, which were eventually implemented into MaizeGDB to find their sequences.

### Chromatin State Mapping of lncRNAs

All the known lncRNAs in maize and their associated chromatin state arrangements using the tool ChromHMM and the chromatin state data from PCSD were mapped. First, a .gtf file that consists of a data list of lncRNAs and *Z. mays* transcript gene IDs (ZM numbers), was obtained from the Nelson lab at Cornell and processed for further analysis. This processing of the list of data, entailed the removal of any duplicate gene IDs, as well as gene ID conversions in which ZM numbers were changed into GRMZM numbers. To remove the duplicates, a python program was used. ChatGPT was used to help write the program (ChatGPT). The program includes a function, which takes the input of the .gtf file for reading and writes an output .gtf file without any transcript gene ID repeats. First the function initialized a set to keep track of the gene IDs.

Then, a for loop iterated through each line of the input .gtf file, which checked and kept track of whether or not the line had a repeated string of characters. If the line had any repeats, it was written to the output once and those without repeats were automatically written to the output file. After the removal of duplicates, the next step for processing the data was to make gene ID conversions. This is because the data from PCSD, used for chromatin state mapping, contains gene IDs with GRMZM format instead of ZM format. To make these conversions, a tool from MaizeGDB known as gene center, which contains libraries of all published types of maize gene ID data, was utilized. GRMZM is an older gene ID format from past databases, while ZM is format for the newer gene IDs in maize. Once the duplicates were removed and converted into GRMZM numbers, batch searches of the data yielded was executed in PCSD, revealing the associated chromatin states of the lncRNAs.

**Using an eQTL Analysis**

From a published database (Fu et al., 2013), eQTL maize data was then obtained and organized into an excel file. Using this eQTL data from *Z. mays,* genes that were targeted for regulation and have association with both lncRNAs and RBPs that cause this regulation, were chosen for analysis. This ultimately revealed the subgroups of the lncRNAs and the RBPs that would be tested for binding.

**Chromatin State Mapping of Target Genes**

All the genes from the eQTL maize data that were targeted for regulation were also mapped using ChromHMM and PCSD. The maize QTL data provided gene IDs for the target genes in GRMZM format, so unlike for the lncRNAs, a conversion was not necessary to proceed. Once all the target genes were mapped, two target genes that had both lncRNAs and RBPs as potential eQTL regulators, were chosen and their chromatins states were analyzed. All the target genes that had the same chromatin states as the two target genes were identified and extracted from the maize eQTL data. This extraction process required a python program, which was made using ChatGPT, in order to identify which target genes had the same chromatin states (ChatGPT). This program reads lines from an input file, processes each line to ensure it contains a specific set of target states, removes duplicate states, and writes the valid lines to an output file. The input file contained all the target genes in maize along with their chromatin states. The program defines a function that reads and processes the data from the input file and stores the information in a list. It then initializes an empty list where the results are stored and another list that has a set of the target chromatin states that each data line must contain in order to be written to the output file. In terms of processing each line, the program used a for loop that iterates through each line of the input file and extracts the data with the states of interest and removes any duplicate states in the data. This yielded two lncRNA subgroups and two Target gene subgroups that would be used to test for binding with the chosen proteins from the maize RBP homologs.

### Predicting lncRNA and RBP Interactions

To test the binding affinities of the maize lncRNAs and RBPs, RPISeq was used. First, two fasta files were generated. The first file contained all the cDNA sequences of the lncRNAs that were in association with the target genes that have the same chromatin states as the first target gene. The second file contained all the cDNA sequences of the lncRNAs that were in association with the target genes that have the same chromatin states as the second target gene. These cDNA sequences were then obtained from MaizeGDB. After generating both fasta files, the maize protein sequences were also generated using MaizeGDB. Finally, RPISeq was used to predict if any binding interactions occurred.

## Results and Discussion

### Certain maize homologs are more similar to human RBPs than others

After using OrthoDB to identify the homologous maize RBPs from the 154 human RBPs available in RBPsuite as CLIP-seq data, 90 maize homologs were found (Figure 1). This finding suggests that 90 of the 154 human RBPs had maize versions like those in humans, indicating that certain maize RBPs share greater similarity with human RBPs than others. For the human RBPs that did not yield maize homologs, several scientific explanations are possible. One explanation is that maize may not possess these RBPs at all, and perhaps never did. Another explanation is that over time, some RBPs in maize and humans could have significantly changed and evolved to a point that OrthoDB does not recognize their sequences as homologs. There could also be functional redundancy where maize have RBPs that perform numerous functions, some like those of the missing homologs, but are still not detectable as homologs in OrthoDB. Additionally, maybe some of the human RBPs are no longer required for maize due to selection or simply because their functions are unnecessary for normal genetic operation in maize. Some of the missing homologs in maize might contain paralogous genes whose functions have significantly diverged due to past gene duplications, making them unrecognizable as homologs to OrthoDB (Kaessmann, 2010). These are some just some of the many plausible scientific reasons for why certain maize homologs exhibit greater similarity to the human RBPs.

### Chromatin mapping of maize lncRNAs revealed that state arrangement is more complex than initially thought

When pursuing this research, one function of lncRNAs that became of interest was their potential influence on regulation through the altering of chromatin structure. In PCSD, there is an extensive amount of data on plant sequences and the chromatin structure of those sequences. Chromatin structure varies throughout the genome, controlling the expression of genes. This variation is classified by the presence of chromatin states. ChromHMM has identified 26 distinct chromatin states in *Z. mays*, each corresponding to different functions and epigenomic marks (Figure 2). The classification of these states is important because it can help predict the regions that either suppress or express genes, allowing the identification of the elements involved. To get an initial idea of how the chromatin states were arranged in the maize transcriptome, the .gtf file that contains the list of lncRNA gene IDs obtained from the Nelson lab, was mapped to their associated chromatin states using ChromHMM and PCSD. By mapping out the maize lncRNAs to their associated chromatin states, it was revealed that the lncRNAs in maize are not associated with a few states, but rather an assorted arrangement of several chromatin states. This information was essential because the original approach was to narrow down the most common chromatin states, identify the lncRNAs associated with those few chromatin states, and see if they bind with RBPs that are also associated with those few chromatin states, but this was not the case. Not only are the 26 maize states, which have numerous marks involved in modifications to the chromatin, arranged in a multitude of combinations, but the frequency for each state also varies. S4 was the most occurring chromatin state, which had 17, 704 hits in the .gtf file of ncRNAs thar contains about 25,000 genes, opposed to S21, which had 47 hits (Figure 3). S4 has a combination of marks but is renowned for having a large presence of histone acetylation, which is usually associated with less condensed chromatin, allowing for the potential expression of genes. S21 is recognized for its regulation properties because it has marks for DNA methylation, which is typically associated with condense and compact chromatin, and Mnase, which can breakdown regions not tightly bound by proteins (Figure2). Overall, if there were fewer states, the chromatin state arrangement would have been much more informative because we might have been able to associate specific states to the range of precise processes that contribute to their presence. Instead, by having multiple diverse arrangements, it can challenge the contextualization of this information and making interpretations based on chromatin structure alone.

**To identify potential regulatory targets of lncRNAs and RBPs, a publicly available eQTL maize dataset was utilized**

Aside from learning about the complex state arrangement in lncRNAs**,** it became evident that if a lncRNA was regulating another locus, the chromatin states of the regulated locus may also be informative. Furthermore, if these targets were to also be regulated by RBPs, then maybe they work with lncRNAs by forming complexes to achieve this. To identify regulatory targets of lncRNAs and RBPs, published eQTL data for maize was used. Two target genes that were identified, TG1 (GRMZM2G108265) and TG2 (GRMZM2G041694), were chosen for having interactions with lncRNAs, L1 (GRMZM2G018006) and L2 (GRMZM2G476477), as well as interactions with binding proteins like maize versions of human RBPs, which were FKBP4, TAF15, and WDR43 (Figure 4). In particular, eQTL was informative in hypothesizing that TG1 is possibly regulated by the interactions between L1 and the two proteins FKBP4 and TAF15, and that TG2 is possibly regulated by the interactions between L2 and WDR43.

**TG1 and TG2 have a complex set of chromatin states with some shared and distinct characteristics**

TG1 has the chromatin states S7, S8, S10, S2, S6, and S4 (Figure 5). 11 target genes were identified and placed in a subgroup for not only having these states, but for also having lncRNA regulators (Figure 6). TG2 has S25, S17, S1, S5, and S4 chromatin states (Figure 5). With 14 target genes that display these states, this subgroup was also identified for being eQTLs of lncRNA regulators (Figure 6). These results can be seen in (Figure 6) which shows the subgroup of target genes from TG1 and their lncRNA regulators (L1), and the subgroup of target genes from TG2 and their lncRNA regulators (L2). Additionally, the figure includes the chromatin states for the target genes from each subgroup (TG1 and TG2). Though all the target genes in the TG1 subgroup have S7, S8, S10, S2, S6, and S4, with frequencies (10,615, 4,191, 3,585, 14,404, 8,287, 17,704), some genes have additional states (Figure 3). This also true for the target genes in the TG2 subgroup, in which some genes have additional states on top of S25, S17, S1, S5, and S4, with frequencies (5,663, 851, 15,266, 14,173, 17,704) (Figure 3). In TG1, the states S4, S6, S7, S8, often share enrichment of histone acetylation, which is often associated with a loss of chromatin compaction and higher levels of transcription, while the states S2 and S10 condense chromatin via methylation, resulting in gene regulation (Figure 2). In TG2, the states S1, S4, and S5, have histone acetylation, while S17 and S25 are enriched in methylation (Figure 2). Additionally, TG1 and TG2 share the most common state, S4 (Figure 3). Another finding I had based on the eQTL analysis, was that some target genes from the two subgroups did not have lncRNAs as regulators. A plausible scientific explanation for this is that this combination of chromatin states can be present with or without lncRNAs as the regulators. Overall, these results highlight the complex arrangement of chromatin states and the distinct functions they induce.

**Using RPISeq to test the affinities for binding interactions between lncRNAs and RBPs in maize yielded output scores that were challenging to interpret**

RPISeq was initially implemented to predict the binding affinities between the lncRNA subgroups (L1 and L2) and the three binding proteins (FKBP4, TAF15, and WDR43). The gene IDs of the lncRNA input for RPISeq are referenced in the right column of (Figure 6), which is a table displaying the target gene subgroups and lncRNA subgroups, along with the chromatin states of the target genes. (Figure 7) shows the protein input for RPISeq, including the maize IDs and sequences of FKBP4, TAF15, and WDR43. The output of RPISeq is represented as prediction scores with values ranging from 0 to 1. Scores above .5 indicate a positive binding prediction but can vary in levels of confidence. Scores closest to 1 are considered high confidence, scores between .5 and .7 indicate moderate confidence, and scores below .5 are categorized as low confidence. RPISeq makes these predictions with two models, an SVM classifier and an RF classifier, each generating their own separate predictions score. Therefore, in the output of RPISeq, each RNA sequence is given two scores.

Based on the maize eQTL, it was hypothesized that L1 and the lncRNAs within its subgroup would have the highest affinities for FKBP4 and TAF15, while L2 and the lncRNAs within its subgroup would have the highest affinities for WDR43. To investigate this, RPISeq was run for both the L1 and L2 subgroups, testing the binding affinities of all lncRNAs within each subgroup against all three RBPs. Specifically, it was anticipated that the prediction scores generated from RPISeq would align with the expectations: high prediction scores for L1 binding with FKBP4 and TAF15 and lower prediction scores for binding with WDR43, and for L2, lower prediction scores for FKBP4 and TAF15 but higher scores for WDR43.

### L1 appeared to moderately bind with all three proteins based on the RPISeq output

When testing the binding affinities between L1 and FKBP4, RPISeq yielded a SVM prediction score of .71 and a RF prediction score of .7 (Figure 8a). This can be seen in the first row of the table in (Figure 8a), which also displays the scores for the other lncRNAs in the L1 subgroup when tested against FKBP4. To determine if the other lncRNAs followed the same trend as L1, a plot was created next to the table in (Figure 8a). This plot is represented as a line graph to visualize all the L1 subgroup scores, showing the trendline of the average SVM and RF values for each lncRNA tested against FKBP4. For each lncRNA, average prediction score was calculated by adding the SVM and RF values and dividing the sum by two. (Figure 10) is a chart I generated that represents this process, except with all lncRNAs scores in L1 and L2 when tested against all three RBPs. Since SVM and RF both find probability scores, I decided to find their average to have scores that represent the entire subgroup. Overall, the plot showed that the scores of the lncRNAs in the L1 subgroup when tested against FKBP4, had values near the prediction scores of L1 (.7 and .71) with the trendline staying within a range of .7 and .75 (Figure 8a). While this doesn't necessarily indicate correlation, it does show that the values of the L1 scores are close to the average prediction scores of the L1 subgroup. Additionally, these scores show that a moderate confidence level of binding for L1 subgroup binding in FKBP4 took place.

When testing the binding affinities between L1 and TAF15, RPISeq yielded a SVM prediction score of .77 and a RF prediction score of .85 (Figure 8b). This can be seen in the first row of the table in (Figure 8b), which also displays the scores for the other lncRNAs in the L1 subgroup when tested against TAF15. To visualize the scores of the other lncRNAs, a line graph showing the trendline of the average SVM and RF values for each lncRNA was created next to the table in (Figure 8b) using the average score values from (Figure 10). Overall, the plot showed that the scores of the lncRNAs in the L1 subgroup when tested for binding with TAF15, stayed near the prediction scores of L1 (.77 and .85) with the trendline ranging from .7 to .8 (Figure 8b). While this doesn't necessarily indicate a strong correlation, it does show that the L1 scores are close to the average prediction scores of the L1 subgroup. Like with FKBP4, these scores indicate a moderate confidence level of binding with L1 subgroup binding in TAF15.

When testing the binding affinities between L1 and WDR43, RPISeq yielded a SVM prediction score of .659 and a RF prediction score of .8 (Figure 8c). This can be seen in the first row of the table in (Figure 8c), which also displays the scores for the other lncRNAs in the L1 subgroup when tested against WDR43. To visualize the scores of the other lncRNAs, a line graph showing the trendline of the average SVM and RF values for each lncRNA was created next to the table in (Figure 8c) using the average score values from (Figure 10). Overall, the plot showed that the scores of the lncRNAs in the L1 subgroup when tested for binding with WDR43, stayed near the prediction scores of L1 (.659 and .8) with the trendline ranging from .65 to .75 (Figure 8c). While this doesn't necessarily indicate a strong correlation, it does show that the L1 scores are close to the average prediction scores of the L1 subgroup. Additionally, the scores indicate a moderate confidence level of binding with L1 subgroup binding in WDR43.

In summary, it was challenging to come to a conclusion about these results. L1 was expected to have higher binding affinities with FKBP4 and TAF15 than with WDR43, but the results showed that it had relatively similar binding affinities with all the proteins. Moreover, the results indicate that there is not only a moderate confidence level binding for all three of the proteins with L1, but for all the lncRNAs in the L1 subgroup as well (Figure 8). Even though the L1 subgroup displayed similar affinities with the three proteins as L1 itself, there is no indication from these results that this was due to chromatin state arrangement. If L1 and its subgroup had an overall lower prediction score for WDR43 binding than for FKBP4 and WDR43, then chromatin structure could have been a plausible explanation for that. Again, this has to do with using maize eQTL data to determine the subgroups of lncRNAs and RBPs. Since L1, FKBP4, and TAF15 were all identified as regulators of the same target gene (TG1), it was reasonable to assume that L1 binds with these two proteins to regulate TG1, specifically with higher confidence than

WDR43 (Figure 4). Additionally, it was plausible to infer that other regulator lncRNAs that have eQTLs of target that exhibit the same chromatin structure and states as TG1, would also bind with FKBP4 and TAF15 (Figure 5). Because this did not happen, a finding from these results is that there isn't a clear explanation as to why the L1 subgroup formed interactions with all three proteins at roughly the same level of confidence.

For all three proteins, L2 appeared to moderately bind but had average subgroup scores slightly higher based on the RPISeq output

When testing the binding affinities between L2 and FKBP4, RPISeq yielded a SVM prediction score of .495 and a RF prediction score of .65 (Figure 9a). This can be seen in the first row of the table in (Figure 9a). To determine if the other lncRNAs followed the same trend as L2, a plot was created next to the table in (Figure 9a). This plot is represented as a line graph to visualize all the L2 subgroup scores, showing the trendline of the average SVM and RF values for each lncRNA tested against FKBP4, obtained from (Figure 10). Overall, the plot showed that the scores of the lncRNAs in the L2 subgroup when tested against FKBP4, had values slightly above the prediction scores of L2 (.495 and .65) with the trendline staying within a range of .65 and .8 (Figure 9a). Despite this, these scores still show that a moderate confidence level of binding for L2 subgroup binding in FKBP4 took place.

When testing the binding affinities between L2 and TAF15, RPISeq yielded a SVM prediction score of .532 and a RF prediction score of .65 (Figure 9b). This can be seen in the first row of the table in (Figure 9b). To visualize the scores of the other lncRNAs, a line graph showing the trendline of the average SVM and RF values for each lncRNA was created next to the table in (Figure 9b) using the average score values from (Figure 10). Overall, the plot showed that the scores of the lncRNAs in the L2 subgroup when tested for binding with TAF15, stayed slightly above the prediction scores of L2 (.532 and .65) with the trendline ranging from .75 to .85 (Figure 9b). Despite this, these scores still indicate a moderate confidence level of binding with L2 subgroup binding in TAF15.

When testing the binding affinities between L2 and WDR43, RPISeq yielded a SVM prediction score of .415 and a RF prediction score of .75 (Figure 9c). This can be seen in the first row of the table in (Figure 9c). To visualize the scores of the other lncRNAs, a line graph showing the trendline of the average SVM and RF values for each lncRNA was created next to the table in (Figure 9c) using the average score values from (Figure 10). Overall, the plot showed that the scores of the lncRNAs in the L2 subgroup when tested for binding with WDR43, stayed slightly above the prediction scores of L1 (.415 and .75) with the trendline ranging from .65 to .8 (Figure 8c). Despite this, the scores indicate a moderate confidence level of binding with L2 subgroup binding in WDR43.

In summary, these results from L2 yielded a similar finding with the L1 results. That being that the RPISeq output is inconclusive. There are no obvious detectable patterns in the data that would support the prediction hypotheses that were made. This is because moderate binding still appears for all three proteins for the L2 subgroup, but L2 itself had lower scores than the average of the subgroup.

**The arrangement of chromatin states varies between the lncRNAs and in the Target genes, but there are detectable patterns that might show why they are arranged in a particular way**

After mapping the chromatin states of both the lncRNAs and target genes, it’s clear that there is a difference in their arrangements (Figure 13). To investigate this difference, I tested the binding affinities of the target genes, TG1 and TG2, with all three proteins, FKBP4, TAF15, and

WDR43, by utilizing RPISeq. The aim was to make state comparisons between lncRNAs and target genes by analyzing high and low binding affinities from each group. I generated (Figure 14), which shows genes with the highest and lowest prediction scores from several categories. The categories include: the highest and lowest scores from L1 and the chromatin state difference between the TG1 subgroup, the highest and lowest scores from TG1 and the chromatin state difference between the L1 subgroup, the highest and lowest scores from L2 and the chromatin state difference between the TG2 subgroup, and the highest and lowest scores from TG2 and the chromatin state difference between the L2 subgroup. The objective was to identify the states shared between the lncRNAs and their associated target genes. Each state has certain epigenetic marks that appear to correspond to both the suppression and expression of genes. LncRNAs that have higher binding affinities influence chromatin states of target genes that suppress genes. LncRNAs that have lower binding affinities, influence chromatin states of target genes that express genes.

### L1 and TG1

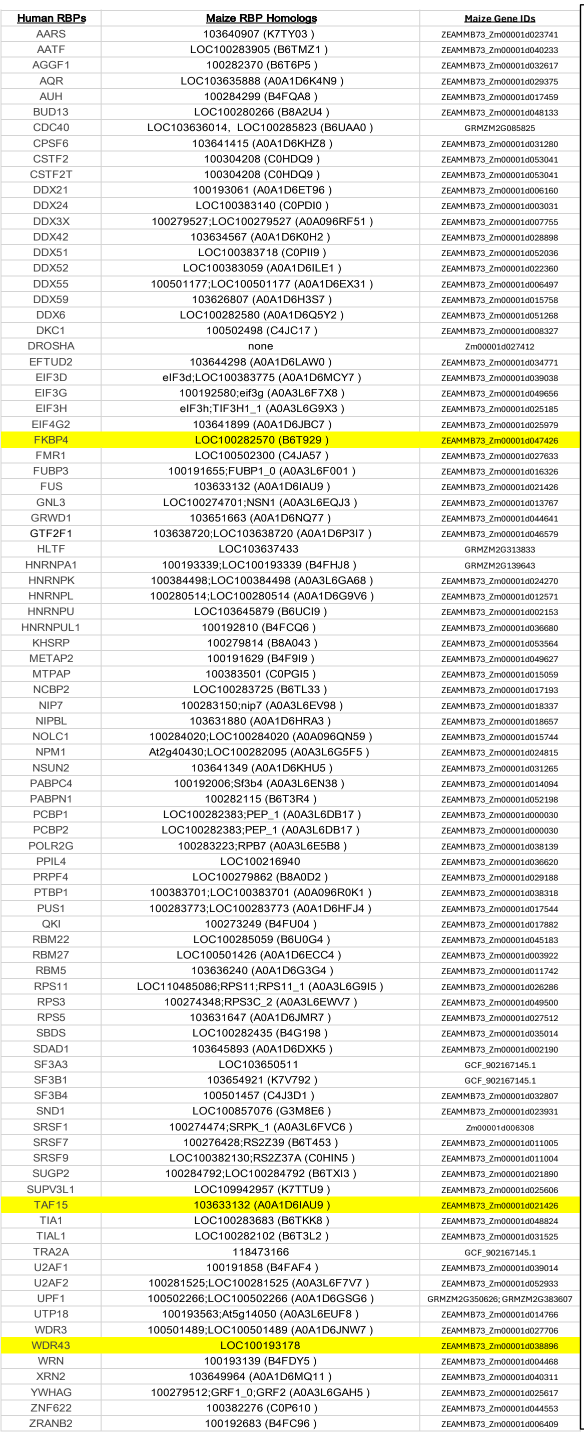
In L1, the chromatin states in the lncRNAs with higher prediction scores have what appears to be a normal amount of methylation and acetylation. The target genes, TG1, predicted to be regulated by the lncRNAs of L1, have states that have normal acetylation, but lots of methylation. For the high L1 scores, the most common states present are S1, S2, S5, and S9. S1, S5, and S9 are noted for their histone acetylation roles, but they also have accessible DNA, an Mnase, and methylation marks that seem normal. S2 is enriched in methylation, and it also has accessible DNA. For the high TG1 scores the states are present in all target genes are S7, S8, S10, S2, S6, S4, along with other common states (S1, S25). The parentheses are to show that S1 and S25 were also present and relatively common but not for all. These states show a normal amount of acetylation and Mnase but now a larger methylation presence. This is significant because the states for the high affinity lncRNAs have epigenetic marks that are the reverse of the states for the target genes, and the high affinity target genes have the epigenetic marks that are the reverse of the lncRNAs. In L1, the chromatin states in the lncRNAs with lower prediction scores have what appears to a lot of acetylation and normal methylation. The target genes in TG1 regulated by these, appear to have a small methylation presence, but an even greater presence of acetylation and states with Mnase marks. For the low L1 scores, the most common states present are S7, S5, S4, S1, which are enriched in acetylation, and S2 renowned for methylation. For the low TG1 scores, S7, S8, S10, S2, S6, and S4 are present for all target genes and are rich in acetylation, having an even greater presence than the low binding lncRNAs in L1. Additional common states (S1, S19, S16, S14, S13, S25), are also present in the low binding TG1 genes, and are also linked to more acetylation, but now large Mnase presence. These results also seem to have the same effect as the high binding groups, in which their chromatin states are reverses of one another. Low binding lncRNAs have chromatin states with more than a normal amount of acetylation, but then influence target genes with states that are more accessible for expression and inhibit regulation, having an even larger presence of acetylation and Mnase presence (Figure 14).

### L2 and TG2

In L2, the chromatin states in the lncRNAs with higher prediction scores have what appears to be a normal amount of methylation and a lot of acetylation. The target genes, TG2, predicted to be regulated by the lncRNAs of L2, have states that have normal acetylation, but lots of methylation and Mnase marks. For the high L2 scores, the most common states present are S1, S2, S9, S5, and S4. S1, S9, S5, and S4 are noted for their histone acetylation roles, but they also have accessible DNA, an Mnase, and methylation marks that seem normal. S2 is enriched in methylation, and it also has accessible DNA. For the high TG2 scores the states are present in all target genes are S25, S17, S1, S5, and S4, which are noted for normal acetylation, methylation, Mnase presence, however, they have the presence of other common states S2, S9, S16, S19, S24. This arrangement of states has huge enrichment of Mnase and methylation, especially in S16, S19, and S24. This is significant because the states for the high affinity lncRNAs have epigenetic marks that are the reverse of the states for the target genes, and the high affinity target genes have the epigenetic marks that are the reverse of the lncRNAs. In L2, the chromatin states in the lncRNAs with lower prediction scores have what appears to be normal acetylation and fewer methylation and Mnase. The target genes in TG2 regulated by these, appear to have a normal methylation presence, but an even greater presence of acetylation and states with Mnase marks. For the low L2 scores, the most common states present are S4, S5, (S7, S9, S1). S4 and S5 are the most common states, while S7, S9, and S1 are still common, just not much. For the low TG2 scores, S25, S17, S1, S5, and S4 are present for all target genes, along with the common states S2 and S7. These results appear to have the same pattern as the L1 and TG1 results, in which lncRNAs from L2 have states with contradicting effects on regulation in relation to TG2 (Figure 14).

# Figures

## Human and Maize RBPs

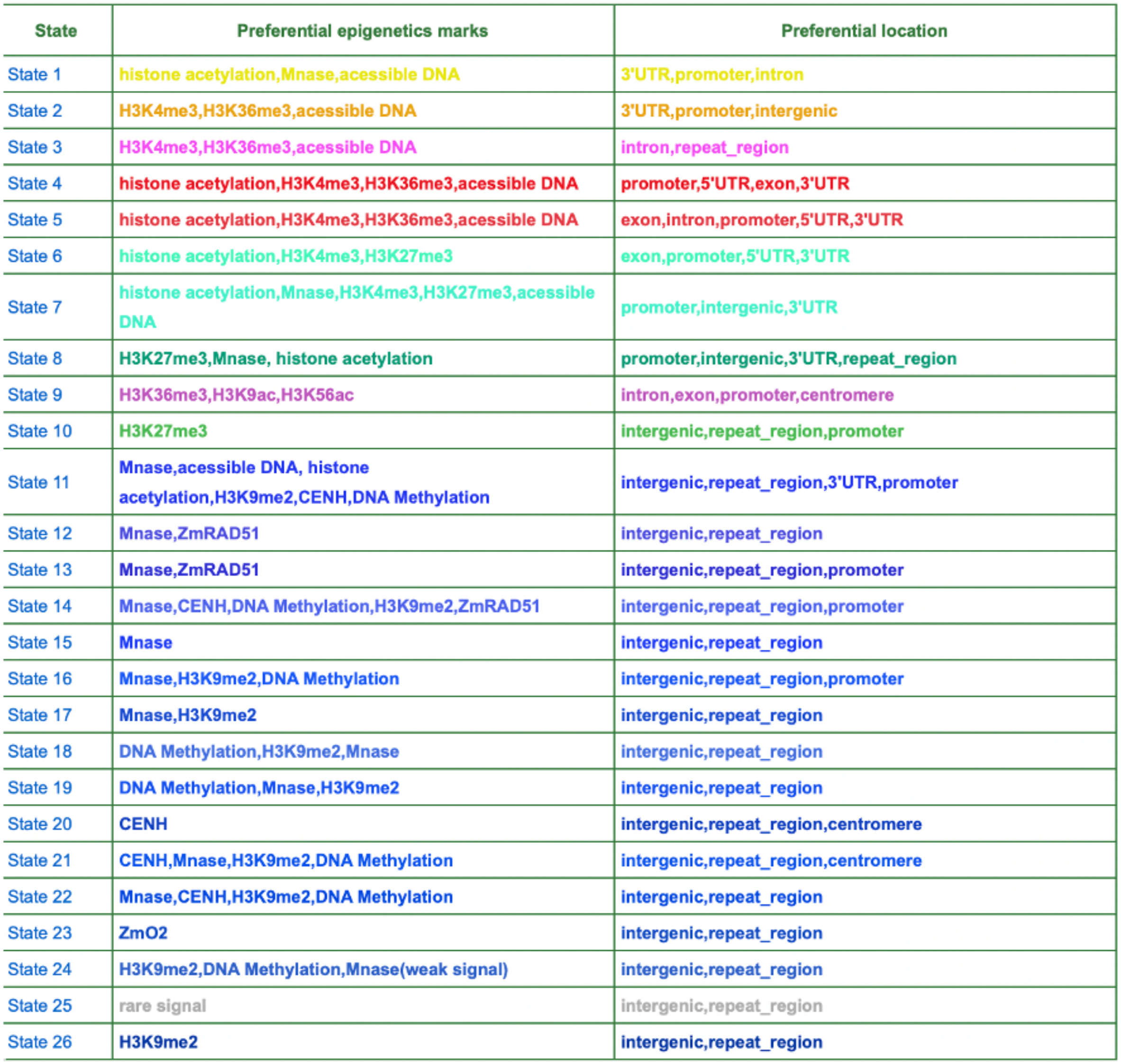


| Human RBPs | Maize RBP Homologs | Maize Gene IDs |
|---|---|---|
| AARS | 103640907 (K7TY03 ) | ZEAMMB73_Zm00001d023741 |
| AATF | LOC100283905 (B6TMZ1 ) | ZEAMMB73_Zm00001d040233 |
| AGGF1 | 100282370 (B6T6P5 ) | ZEAMMB73_Zm00001d032617 |
| AQR | LOC103635888 (A0A1D6K4N9 ) | ZEAMMB73_Zm00001d029375 |
| AUH | 100284299 (B4FQA8 ) | ZEAMMB73_Zm00001d017459 |
| BUD13 | LOC100280266 (B8A2U4 ) | ZEAMMB73_Zm00001d048133 |
| CDC40 | LOC103636014,  LOC100285823 (B6UAA0 ) | GRMZM2G085825 |
| CPSF6 | 103641415 (A0A1D6KHZ8 ) | ZEAMMB73_Zm00001d031280 |
| CSTF2 | 100304208 (C0HDQ9 ) | ZEAMMB73_Zm00001d053041 |
| CSTF2T | 100304208 (C0HDQ9 ) | ZEAMMB73_Zm00001d053041 |
| DDX21 | 100193061 (A0A1D6ET96 ) | ZEAMMB73_Zm00001d006160 |
| DDX24 | LOC100383140 (C0PDI0 ) | ZEAMMB73_Zm00001d003031 |
| DDX3X | 100279527;LOC100279527 (A0A096RF51 ) | ZEAMMB73_Zm00001d007755 |
| DDX42 | 103634567 (A0A1D6K0H2 ) | ZEAMMB73_Zm00001d028898 |
| DDX51 | LOC100383718 (C0PII9 ) | ZEAMMB73_Zm00001d052036 |
| DDX52 | LOC100383059 (A0A1D6ILE1 ) | ZEAMMB73_Zm00001d022360 |
| DDX55 | 100501177;LOC100501177 (A0A1D6EX31 ) | ZEAMMB73_Zm00001d006497 |
| DDX59 | 103626807 (A0A1D6H3S7 ) | ZEAMMB73_Zm00001d015758 |
| DDX6 | LOC100282580 (A0A1D6Q5Y2 ) | ZEAMMB73_Zm00001d051268 |
| DKC1 | 100502498 (C4JC17 ) | ZEAMMB73_Zm00001d008327 |
| DROSHA | none | Zm00001d027412 |
| EFTUD2 | 103644298 (A0A1D6LAW0 ) | ZEAMMB73_Zm00001d034771 |
| EIF3D | eIF3d;LOC100383775 (A0A1D6MCY7 ) | ZEAMMB73_Zm00001d039038 |
| EIF3G | 100192580;eif3g (A0A3L6F7X8 ) | ZEAMMB73_Zm00001d049656 |
| EIF3H | eIF3h;TIF3H1_1 (A0A3L6G9X3 ) | ZEAMMB73_Zm00001d025185 |
| EIF4G2 | 103641899 (A0A1D6JBC7 ) | ZEAMMB73_Zm00001d025979 |
| FKBP4 | LOC100282570 (B6T929 ) | ZEAMMB73_Zm00001d047426 |
| FMR1 | LOC100502300 (C4JA57 ) | ZEAMMB73_Zm00001d027633 |
| FUBP3 | 100191655;FUBP1_0 (A0A3L6F001 ) | ZEAMMB73_Zm00001d016326 |
| FUS | 103633132 (A0A1D6IAU9 ) | ZEAMMB73_Zm00001d021426 |
| GNL3 | LOC100274701;NSN1 (A0A3L6EQJ3 ) | ZEAMMB73_Zm00001d013767 |
| GRWD1 | 103651663 (A0A1D6NQ77 ) | ZEAMMB73_Zm00001d044641 |
| GTF2F1 | 103638720;LOC103638720 (A0A1D6P3I7 ) | ZEAMMB73_Zm00001d046579 |
| HLTF | LOC103637433 | GRMZM2G313833 |
| HNRNPA1 | 100193339;LOC100193339 (B4FHJ8 ) | GRMZM2G139643 |
| HNRNPK | 100384498;LOC100384498 (A0A3L6GA68 ) | ZEAMMB73_Zm00001d024270 |
| HNRNPL | 100280514;LOC100280514 (A0A1D6G9V6 ) | ZEAMMB73_Zm00001d012571 |
| HNRNPU | LOC103645879 (B6UCI9 ) | ZEAMMB73_Zm00001d002153 |
| HNRNPUL1 | 100192810 (B4FCQ6 ) | ZEAMMB73_Zm00001d036680 |
| KHSRP | 100279814 (B8A043 ) | ZEAMMB73_Zm00001d053564 |
| METAP2 | 100191629 (B4F9I9 ) | ZEAMMB73_Zm00001d049627 |
| MTPAP | 100383501 (C0PGI5 ) | ZEAMMB73_Zm00001d015059 |
| NCBP2 | LOC100283725 (B6TL33 ) | ZEAMMB73_Zm00001d017193 |
| NIP7 | 100283150;nip7 (A0A3L6EV98 ) | ZEAMMB73_Zm00001d018337 |
| NIPBL | 103631880 (A0A1D6HRA3 ) | ZEAMMB73_Zm00001d018657 |
| NOLC1 | 100284020;LOC100284020 (A0A096QN59 ) | ZEAMMB73_Zm00001d015744 |
| NPM1 | At2g40430;LOC100282095 (A0A3L6G5F5 ) | ZEAMMB73_Zm00001d024815 |
| NSUN2 | 103641349 (A0A1D6KHU5 ) | ZEAMMB73_Zm00001d031265 |
| PABPC4 | 100192006;Sf3b4 (A0A3L6EN38 ) | ZEAMMB73_Zm00001d014094 |
| PABPN1 | 100282115 (B6T3R4 ) | ZEAMMB73_Zm00001d052198 |
| PCBP1 | LOC100282383;PEP_1 (A0A3L6DB17 ) | ZEAMMB73_Zm00001d000030 |
| PCBP2 | LOC100282383;PEP_1 (A0A3L6DB17 ) | ZEAMMB73_Zm00001d000030 |
| POLR2G | 100283223;RPB7 (A0A3L6E5B8 ) | ZEAMMB73_Zm00001d038139 |
| PPIL4 | LOC100216940 | ZEAMMB73_Zm00001d036620 |
| PRPF4 | LOC100279862 (B8A0D2 ) | ZEAMMB73_Zm00001d029188 |
| PTBP1 | 100383701;LOC100383701 (A0A096R0K1 ) | ZEAMMB73_Zm00001d038318 |
| PUS1 | 100283773;LOC100283773 (A0A1D6HFJ4 ) | ZEAMMB73_Zm00001d017544 |
| QKI | 100273249 (B4FU04 ) | ZEAMMB73_Zm00001d017882 |
| RBM22 | LOC100285059 (B6U0G4 ) | ZEAMMB73_Zm00001d045183 |
| RBM27 | LOC100501426 (A0A1D6ECC4 ) | ZEAMMB73_Zm00001d003922 |
| RBM5 | 103636240 (A0A1D6G3G4 ) | ZEAMMB73_Zm00001d011742 |
| RPS11 | LOC110485086;RPS11;RPS11_1 (A0A3L6G9I5 ) | ZEAMMB73_Zm00001d026286 |
| RPS3 | 100274348;RPS3C_2 (A0A3L6EWV7 ) | ZEAMMB73_Zm00001d049500 |
| RPS5 | 103631647 (A0A1D6JMR7 ) | ZEAMMB73_Zm00001d027512 |
| SBDS | LOC100282435 (B4G198 ) | ZEAMMB73_Zm00001d035014 |
| SDAD1 | 103645893 (A0A1D6DXK5 ) | ZEAMMB73_Zm00001d002190 |
| SF3A3 | LOC103650511 | GCF_902167145.1 |
| SF3B1 | 103654921 (K7V792 ) | GCF_902167145.1 |
| SF3B4 | 100501457 (C4J3D1 ) | ZEAMMB73_Zm00001d032807 |
| SND1 | LOC100857076 (G3M8E6 ) | ZEAMMB73_Zm00001d023931 |
| SRSF1 | 100274474;SRPK_1 (A0A3L6FVC6 ) | Zm00001d006308 |
| SRSF7 | 100276428;RS2Z39 (B6T453 ) | ZEAMMB73_Zm00001d011005 |
| SRSF9 | LOC100382130;RS2Z37A (C0HIN5 ) | ZEAMMB73_Zm00001d011004 |
| SUGP2 | 100284792;LOC100284792 (B6TXI3 ) | ZEAMMB73_Zm00001d021890 |
| SUPV3L1 | LOC109942957 (K7TTU9 ) | ZEAMMB73_Zm00001d025606 |
| TAF15 | 103633132 (A0A1D6IAU9 ) | ZEAMMB73_Zm00001d021426 |
| TIA1 | LOC100283683 (B6TKK8 ) | ZEAMMB73_Zm00001d048824 |
| TIAL1 | LOC100282102 (B6T3L2 ) | ZEAMMB73_Zm00001d031525 |
| TRA2A | 118473166 | GCF_902167145.1 |
| U2AF1 | 100191858 (B4FAF4 ) | ZEAMMB73_Zm00001d039014 |
| U2AF2 | 100281525;LOC100281525 (A0A3L6F7V7 ) | ZEAMMB73_Zm00001d052933 |
| UPF1 | 100502266;LOC100502266 (A0A1D6GSG6 ) | GRMZM2G350626; GRMZM2G383607 |
| UTP18 | 100193563;At5g14050 (A0A3L6EUF8 ) | ZEAMMB73_Zm00001d014766 |
| WDR3 | 100501489;LOC100501489 (A0A1D6JNW7 ) | ZEAMMB73_Zm00001d027706 |
| WDR43 | LOC100193178 | ZEAMMB73_Zm00001d038896 |
| WRN | 100193139 (B4FDY5 ) | ZEAMMB73_Zm00001d004468 |
| XRN2 | 103649964 (A0A1D6MQ11 ) | ZEAMMB73_Zm00001d040311 |
| YWHAG | 100279512;GRF1_0;GRF2 (A0A3L6GAH5 ) | ZEAMMB73_Zm00001d025617 |
| ZNF622 | 100382276 (C0P610 ) | ZEAMMB73_Zm00001d044553 |
| ZRANB2 | 100192683 (B4FC96 ) | ZEAMMB73_Zm00001d006409 |

Figure 1

This table is the result of implementing the CLIP-seq data of 154 human RBPs into OrthoDB to find the maize RBPs. This yielded 91 maize orthologs which was then searched in MaizeGDB, revealing their associated gene IDs. The left column contains the human RBP names. The middle column has the maize orthologs. The right column has the maize gene IDs. The highlighted rows are the proteins used in RPISeq.

## Maize Chromatin States

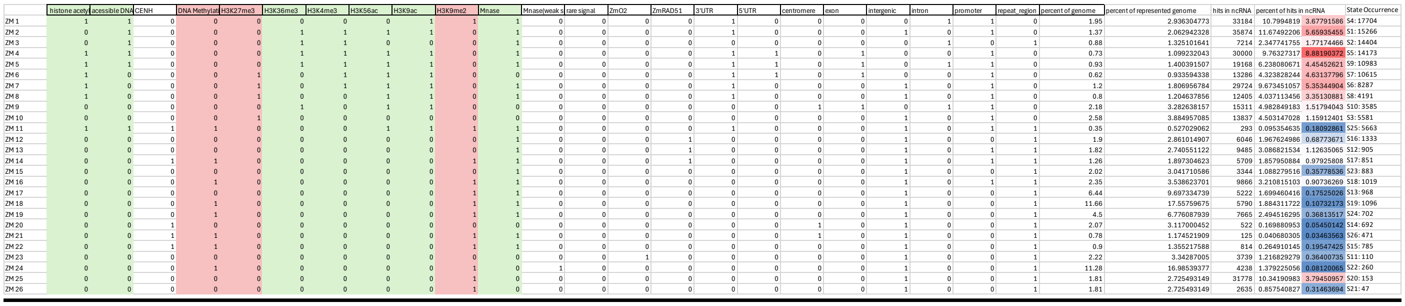


| State | Preferential epigenetics marks | Preferential location |
|---|---|---|
| State 1 | histone acetylation,Mnase,acessible DNA | 3'UTR,promoter,intron |
| State 2 | H3K4me3,H3K36me3,acessible DNA | 3'UTR,promoter,intergenic |
| State 3 | H3K4me3,H3K36me3,acessible DNA | intron,repeat_region |
| State 4 | histone acetylation,H3K4me3,H3K36me3,acessible DNA | promoter,5'UTR,exon,3'UTR |
| State 5 | histone acetylation,H3K4me3,H3K36me3,acessible DNA | exon,intron,promoter,5'UTR,3'UTR |
| State 6 | histone acetylation,H3K4me3,H3K27me3 | exon,promoter,5'UTR,3'UTR |
| State 7 | histone acetylation,Mnase,H3K4me3,H3K27me3,acessible DNA | promoter,intergenic,3'UTR |
| State 8 | H3K27me3,Mnase, histone acetylation | promoter,intergenic,3'UTR,repeat_region |
| State 9 | H3K36me3,H3K9ac,H3K56ac | intron,exon,promoter,centromere |
| State 10 | H3K27me3 | intergenic,repeat_region,promoter |
| State 11 | Mnase,acessible DNA, histone acetylation,H3K9me2,CENH,DNA Methylation | intergenic,repeat_region,3'UTR,promoter |
| State 12 | Mnase,ZmRAD51 | intergenic,repeat_region |
| State 13 | Mnase,ZmRAD51 | intergenic,repeat_region,promoter |
| State 14 | Mnase,CENH,DNA Methylation,H3K9me2,ZmRAD51 | intergenic,repeat_region,promoter |
| State 15 | Mnase | intergenic,repeat_region |
| State 16 | Mnase,H3K9me2,DNA Methylation | intergenic,repeat_region,promoter |
| State 17 | Mnase,H3K9me2 | intergenic,repeat_region |
| State 18 | DNA Methylation,H3K9me2,Mnase | intergenic,repeat_region |
| State 19 | DNA Methylation,Mnase,H3K9me2 | intergenic,repeat_region |
| State 20 | CENH | intergenic,repeat_region,centromere |
| State 21 | CENH,Mnase,H3K9me2,DNA Methylation | intergenic,repeat_region,centromere |
| State 22 | Mnase,CENH,H3K9me2,DNA Methylation | intergenic,repeat_region |
| State 23 | ZmO2 | intergenic,repeat_region |
| State 24 | H3K9me2,DNA Methylation,Mnase(weak signal) | intergenic,repeat_region |
| State 25 | rare signal | intergenic,repeat_region |
| State 26 | H3K9me2 | intergenic,repeat_region |

Figure 2

This chart was obtained from PCSD, and it shows all 26 states found in maize according to ChromHMM along with their epigenetic marks and preferential location (PCSD).

## Maize Chromatin Marks and Stats



| | histone acety | acessible DNA | CENH | DNA Methylati | H3K27me3 | H3K36me3 | H3K4me3 | H3K56ac | H3K9ac | H3K9me2 | Mnase | Mnase(weak s | rare signal | ZmO2 | ZmRAD51 | 3'UTR | 5'UTR | centromere | exon | intergenic | intron | promoter | repeat_region | percent of genome | percent of represented genome | hits in ncRNA | percent of hits in ncRNA | | State Occurrence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZM 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1.95 | 2.936304773 | 33184 | 10.7994819 | 3.67791586 | S4: 17704 |
| ZM 2 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1.37 | 2.062942328 | 35874 | 11.67492206 | 5.65935455 | S1: 15266 |
| ZM 3 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.88 | 1.325101641 | 7214 | 2.347741755 | 1.77174466 | S2: 14404 |
| ZM 4 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0.73 | 1.099232043 | 30000 | 9.76327317 | 8.88190372 | S5: 14173 |
| ZM 5 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0.93 | 1.400391507 | 19168 | 6.238080671 | 4.45452621 | S9: 10983 |
| ZM 6 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0.62 | 0.933594338 | 13286 | 4.323828244 | 4.63137796 | S7: 10615 |
| ZM 7 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1.2 | 1.806956784 | 29724 | 9.673451057 | 5.35344904 | S6: 8287 |
| ZM 8 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0.8 | 1.204637856 | 12405 | 4.037113456 | 3.35130881 | S8: 4191 |
| ZM 9 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 2.18 | 3.282638157 | 15311 | 4.982849183 | 1.51794043 | S10: 3585 |
| ZM 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 2.58 | 3.884957085 | 13837 | 4.503147028 | 1.15912401 | S3: 5581 |
| ZM 11 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0.35 | 0.527029062 | 293 | 0.095354635 | 0.18092861 | S25: 5663 |
| ZM 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1.9 | 2.861014907 | 6046 | 1.967624986 | 0.68773671 | S16: 1333 |
| ZM 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1.82 | 2.740551122 | 9485 | 3.086821534 | 1.12635065 | S12: 905 |
| ZM 14 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1.26 | 1.897304623 | 5709 | 1.857950884 | 0.97925808 | S17: 851 |
| ZM 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2.02 | 3.041710586 | 3344 | 1.088279516 | 0.35778536 | S23: 883 |
| ZM 16 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2.35 | 3.538623701 | 9866 | 3.210815103 | 0.90736269 | S18: 1019 |
| ZM 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 6.44 | 9.697334739 | 5222 | 1.699460416 | 0.17525026 | S13: 968 |
| ZM 18 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 11.66 | 17.55759675 | 5790 | 1.884311722 | 0.10732173 | S19: 1096 |
| ZM 19 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 4.5 | 6.776087939 | 7665 | 2.494516295 | 0.36813517 | S24: 702 |
| ZM 20 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 2.07 | 3.117000452 | 522 | 0.169880953 | 0.05450142 | S14: 692 |
| ZM 21 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0.78 | 1.174521909 | 125 | 0.040680305 | 0.03463563 | S26: 471 |
| ZM 22 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0.9 | 1.355217588 | 814 | 0.264910145 | 0.19547425 | S15: 785 |
| ZM 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2.22 | 3.34287005 | 3739 | 1.216829279 | 0.36400735 | S11: 110 |
| ZM 24 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 11.28 | 16.98539377 | 4238 | 1.379225056 | 0.08120065 | S22: 260 |
| ZM 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1.81 | 2.725493149 | 31778 | 10.34190983 | 3.79450957 | S20: 153 |
| ZM 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1.81 | 2.725493149 | 2635 | 0.857540827 | 0.31463694 | S21: 47 |

Figure 3

This table shows all 26 chromatin states in maize along with the epigenetic marks identified with ChromHMM. A value of 1 means the mark is present and a value of 0 means it is not present. The green columns represent marks that typically yield gene expression while the red columns represent marks that typically repress gene expression. The 5 columns on the right side of the table are some statistics that helped to quantify the data. These include the percent each state is represented in the genome, the exact number and percent of ncRNA hits in the maize genome, and the occurrence of each state.

## eQTL Target Genes and Regulators

| **Regulatory Target Genes** | **lncRNA Regulators** | **RBP Regulators** |
|---|---|---|
| TG1 (GRMZM2G108265) | L1 (GRMZM2G018006) | FKBP4<br>TAF15 |
| TG2 (GRMZM2G041694) | L2 (GRMZM2G476477) | WDR43 |

Figure 4

This tables shows the two target genes (TG1 and TG2) obtained from the maize eQTL data and their chromatin states. The middle column has the two lncRNAs (L1 and L2) that are eQTLs of the target genes. The column to the right represents the RBPs that were also predicted to be regulators. TG1 is predicted to be regulated by L1, FKBP4, and TAF15. TG2 is predicted to be regulated by L2 and WDR43.

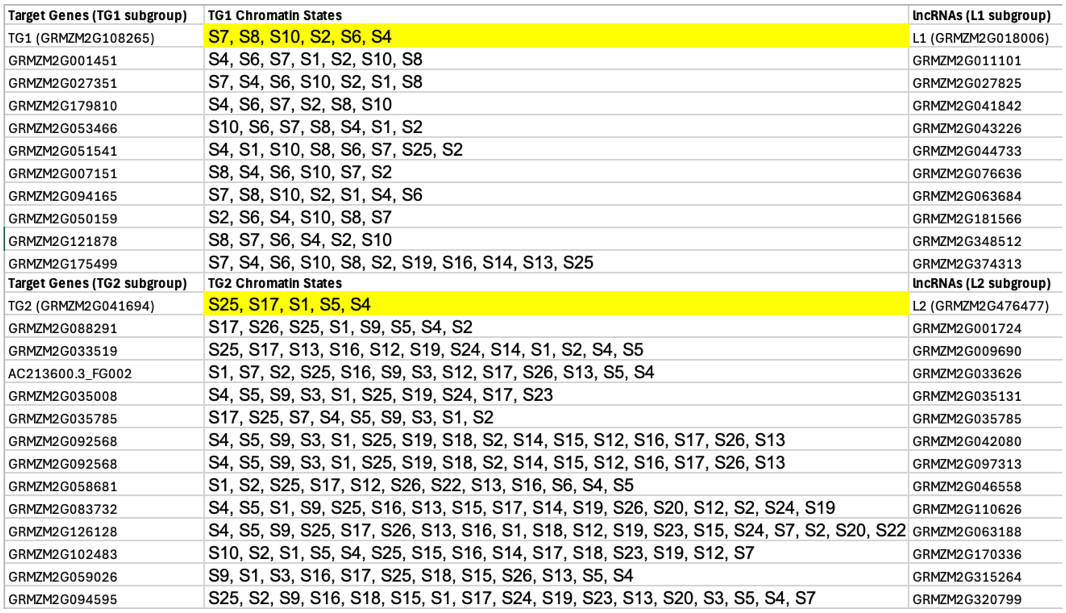
### Target Gene States

| Regulatory Target Genes | TG Chromatin States |
|---|---|
| TG1 (GRMZM2G108265) | S7, S8, S10, S2, S6, S4 |
| TG2 (GRMZM2G041694) | S25, S17, S1, S5, S4 |

Figure 5

The left column contains the regulatory target genes, TG1 and TG2, while the right column are their associated chromatin states, which were obtained using ChromHMM.

### Subgroup of Target Genes with States and lncRNA Regulators

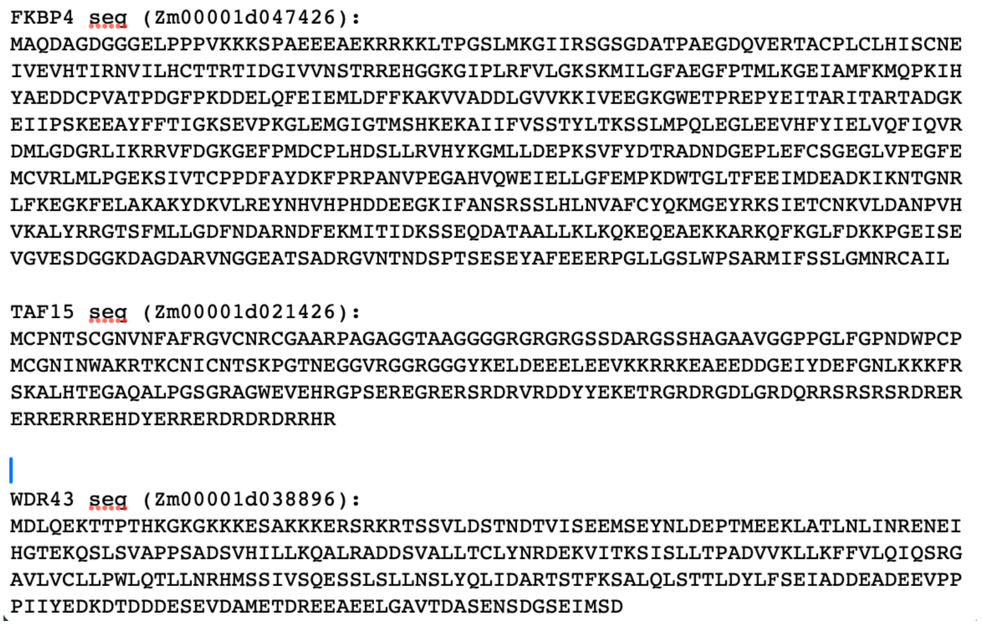


| Target Genes (TG1 subgroup) | TG1 Chromatin States | lncRNAs (L1 subgroup) |
|---|---|---|
| TG1 (GRMZM2G108265) | S7, S8, S10, S2, S6, S4 | L1 (GRMZM2G018006) |
| GRMZM2G001451 | S4, S6, S7, S1, S2, S10, S8 | GRMZM2G011101 |
| GRMZM2G027351 | S7, S4, S6, S10, S2, S1, S8 | GRMZM2G027825 |
| GRMZM2G179810 | S4, S6, S7, S2, S8, S10 | GRMZM2G041842 |
| GRMZM2G053466 | S10, S6, S7, S8, S4, S1, S2 | GRMZM2G043226 |
| GRMZM2G051541 | S4, S1, S10, S8, S6, S7, S25, S2 | GRMZM2G044733 |
| GRMZM2G007151 | S8, S4, S6, S10, S7, S2 | GRMZM2G076636 |
| GRMZM2G094165 | S7, S8, S10, S2, S1, S4, S6 | GRMZM2G063684 |
| GRMZM2G050159 | S2, S6, S4, S10, S8, S7 | GRMZM2G181566 |
| GRMZM2G121878 | S8, S7, S6, S4, S2, S10 | GRMZM2G348512 |
| GRMZM2G175499 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 | GRMZM2G374313 |
| **Target Genes (TG2 subgroup)** | **TG2 Chromatin States** | **lncRNAs (L2 subgroup)** |
| TG2 (GRMZM2G041694) | S25, S17, S1, S5, S4 | L2 (GRMZM2G476477) |
| GRMZM2G088291 | S17, S26, S25, S1, S9, S5, S4, S2 | GRMZM2G001724 |
| GRMZM2G033519 | S25, S17, S13, S16, S12, S19, S24, S14, S1, S2, S4, S5 | GRMZM2G009690 |
| AC213600.3_FG002 | S1, S7, S2, S25, S16, S9, S3, S12, S17, S26, S13, S5, S4 | GRMZM2G033626 |
| GRMZM2G035008 | S4, S5, S9, S3, S1, S25, S19, S24, S17, S23 | GRMZM2G035131 |
| GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | GRMZM2G035785 |
| GRMZM2G092568 | S4, S5, S9, S3, S1, S25, S19, S18, S2, S14, S15, S12, S16, S17, S26, S13 | GRMZM2G042080 |
| GRMZM2G092568 | S4, S5, S9, S3, S1, S25, S19, S18, S2, S14, S15, S12, S16, S17, S26, S13 | GRMZM2G097313 |
| GRMZM2G058681 | S1, S2, S25, S17, S12, S26, S22, S13, S16, S6, S4, S5 | GRMZM2G046558 |
| GRMZM2G083732 | S4, S5, S1, S9, S25, S16, S13, S15, S17, S14, S19, S26, S20, S12, S2, S24, S19 | GRMZM2G110626 |
| GRMZM2G126128 | S4, S5, S9, S25, S17, S26, S13, S16, S1, S18, S12, S19, S23, S15, S24, S7, S2, S20, S22 | GRMZM2G063188 |
| GRMZM2G102483 | S10, S2, S1, S5, S4, S25, S15, S16, S14, S17, S18, S23, S19, S12, S7 | GRMZM2G170336 |
| GRMZM2G059026 | S9, S1, S3, S16, S17, S25, S18, S15, S26, S13, S5, S4 | GRMZM2G315264 |
| GRMZM2G094595 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 | GRMZM2G320799 |

Figure 6

This figure shows the 2 target genes obtained from the eQTL data along with their chromatin state arrangements. The highlighted sections are the 2 target genes, below them in the first column are the target genes that have the same states as them. To the right in the third column are the gene IDs of the lncRNAs that were considered regulators in the eQTL data. The sequences of these IDs were used in RPISeq. This is the resulting subgroup of lncRNAs used in RPISeq to predict any binding interactions with maize versions of FKBP4, TAF15, and WDR43.

## Input Protein Sequences for RPISeq



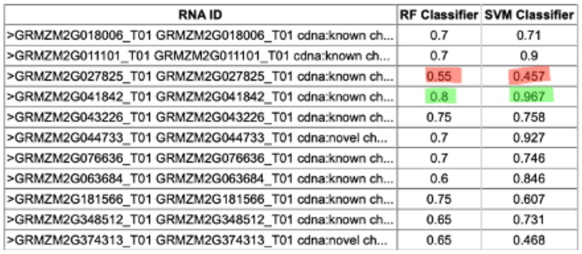
```
FKBP4 seq (Zm00001d047426):
MAQDAGDGGGELPPPVKKKSPAEEEAEKRRKKLTPGSLMKGIIRSGSGDATPAEGDQVERTACPLCLHISCNE
IVEVHTIRNVILHCTTRTIDGIVVNSTRREHGGKGIPLRFVLGKSKMILGFAEGFPTMLKGEIAMFKMQPKIH
YAEDDCPVATPDGFPKDDELQFEIEMLDFFKAKVVADDLGVVKKIVEEGKGWETPREPYEITARITARTADGK
EIIPSKEEAYFFTIGKSEVPKGLEMGIGTMSHKEKAIIFVSSTYLTKSSLMPQLEGLEEVHFYIELVQFIQVR
DMLGDGRLIKRRVFDGKGEFPMDCPLHDSLLRVHYKGMLLDEPKSVFYDTRADNDGEPLEFCSGEGLVPEGFE
MCVRLMLPGEKSIVTCPPDFAYDKFPRPANVPEGAHVQWEIELLGFEMPKDWTGLTFEEIMDEADKIKNTGNR
LFKEGKFELAKAKYDKVLREYNHVHPHDDEEGKIFANSRSSLHLNVAFCYQKMGEYRKSIETCNKVLDANPVH
VKALYRRGTSFMLLGDFNDARNDFEKMITIDKSSEQDATAALLKLKQKEQEAEKKARKQFKGLFDKKPGEISE
VGVESDGGKDAGDARVNGGEATSADRGVNTNDSPTSESEYAFEEERPGLLGSLWPSARMIFSSLGMNRCAIL

TAF15 seq (Zm00001d021426):
MCPNTSCGNVNFAFRGVCNRCGAARPAGAGGTAAGGGGRGRGRGSSDARGSSHAGAAVGGPPGLFGPNDWPCP
MCGNINWAKRTKCNICNTSKPGTNEGGVRGGRGGGYKELDEEELEEVKKRRKEAEEDDGEIYDEFGNLKKKFR
SKALHTEGAQALPGSGRAGWEVEHRGPSEREGRERSRDRVRDDYYEKETRGRDRGDLGRDQRRSRSRSRDRER
ERRERRREHDYERRERDRDRDRRHR

WDR43 seq (Zm00001d038896):
MDLQEKTTPTHKGKGKKKESAKKKERSRKRTSSVLDSTNDTVISEEMSEYNLDEPTMEEKLATLNLINRENEI
HGTEKQSLSVAPPSADSVHILLKQALRADDSVALLTCLYNRDEKVITKSISLLTPADVVKLLKFFVLQIQSRG
AVLVCLLPWLQTLLNRHMSSIVSQESSLSLLNSLYQLIDARTSTFKSALQLSTTLDYLFSEIADDEADEEVPP
PIIYEDKDTDDDESEVDAMETDREEAEELGAVTDASENSDGSEIMSD
```

Figure 7

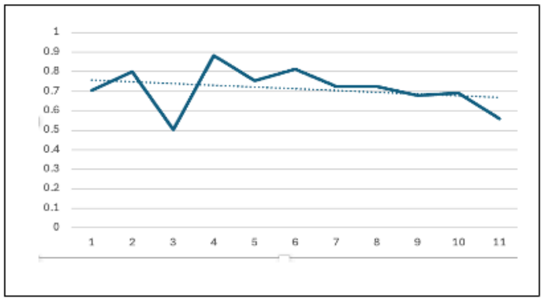
This figure shows the sequences and gene IDs of the proteins used in RPISeq. The gene IDs were implemented in MaizeGDB in order to get these sequences, resulting in the above translation sequences.

**RPISeq Predictions for L1**

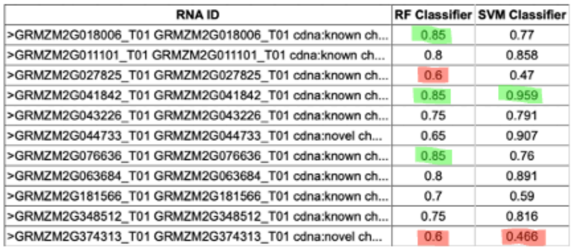
a. FKBP4/L1



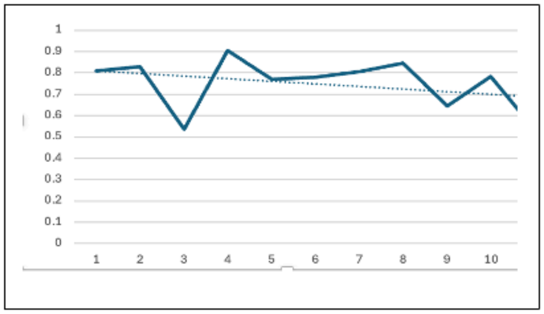
| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G018006_T01 GRMZM2G018006_T01 cdna:known ch... | 0.7 | 0.71 |
| >GRMZM2G011101_T01 GRMZM2G011101_T01 cdna:known ch... | 0.7 | 0.9 |
| >GRMZM2G027825_T01 GRMZM2G027825_T01 cdna:known ch... | 0.55 | 0.457 |
| >GRMZM2G041842_T01 GRMZM2G041842_T01 cdna:known ch... | 0.8 | 0.967 |
| >GRMZM2G043226_T01 GRMZM2G043226_T01 cdna:known ch... | 0.75 | 0.758 |
| >GRMZM2G044733_T01 GRMZM2G044733_T01 cdna:novel ch... | 0.7 | 0.927 |
| >GRMZM2G076636_T01 GRMZM2G076636_T01 cdna:known ch... | 0.7 | 0.746 |
| >GRMZM2G063684_T01 GRMZM2G063684_T01 cdna:known ch... | 0.6 | 0.846 |
| >GRMZM2G181566_T01 GRMZM2G181566_T01 cdna:known ch... | 0.75 | 0.607 |
| >GRMZM2G348512_T01 GRMZM2G348512_T01 cdna:known ch... | 0.65 | 0.731 |
| >GRMZM2G374313_T01 GRMZM2G374313_T01 cdna:novel ch... | 0.65 | 0.468 |



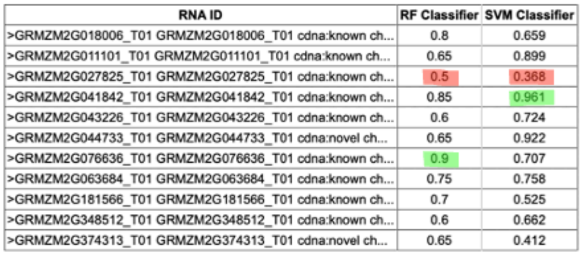
b. TAF15/L1



| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G018006_T01 GRMZM2G018006_T01 cdna:known ch... | 0.85 | 0.77 |
| >GRMZM2G011101_T01 GRMZM2G011101_T01 cdna:known ch... | 0.8 | 0.858 |
| >GRMZM2G027825_T01 GRMZM2G027825_T01 cdna:known ch... | 0.6 | 0.47 |
| >GRMZM2G041842_T01 GRMZM2G041842_T01 cdna:known ch... | 0.85 | 0.959 |
| >GRMZM2G043226_T01 GRMZM2G043226_T01 cdna:known ch... | 0.75 | 0.791 |
| >GRMZM2G044733_T01 GRMZM2G044733_T01 cdna:novel ch... | 0.65 | 0.907 |
| >GRMZM2G076636_T01 GRMZM2G076636_T01 cdna:known ch... | 0.85 | 0.76 |
| >GRMZM2G063684_T01 GRMZM2G063684_T01 cdna:known ch... | 0.8 | 0.891 |
| >GRMZM2G181566_T01 GRMZM2G181566_T01 cdna:known ch... | 0.7 | 0.59 |
| >GRMZM2G348512_T01 GRMZM2G348512_T01 cdna:known ch... | 0.75 | 0.816 |
| >GRMZM2G374313_T01 GRMZM2G374313_T01 cdna:novel ch... | 0.6 | 0.466 |



c. WDR43/L1



| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G018006_T01 GRMZM2G018006_T01 cdna:known ch... | 0.8 | 0.659 |
| >GRMZM2G011101_T01 GRMZM2G011101_T01 cdna:known ch... | 0.65 | 0.899 |
| >GRMZM2G027825_T01 GRMZM2G027825_T01 cdna:known ch... | 0.5 | 0.368 |
| >GRMZM2G041842_T01 GRMZM2G041842_T01 cdna:known ch... | 0.85 | 0.961 |
| >GRMZM2G043226_T01 GRMZM2G043226_T01 cdna:known ch... | 0.6 | 0.724 |
| >GRMZM2G044733_T01 GRMZM2G044733_T01 cdna:novel ch... | 0.65 | 0.922 |
| >GRMZM2G076636_T01 GRMZM2G076636_T01 cdna:known ch... | 0.9 | 0.707 |
| >GRMZM2G063684_T01 GRMZM2G063684_T01 cdna:known ch... | 0.75 | 0.758 |
| >GRMZM2G181566_T01 GRMZM2G181566_T01 cdna:known ch... | 0.7 | 0.525 |
| >GRMZM2G348512_T01 GRMZM2G348512_T01 cdna:known ch... | 0.6 | 0.662 |
| >GRMZM2G374313_T01 GRMZM2G374313_T01 cdna:novel ch... | 0.65 | 0.412 |

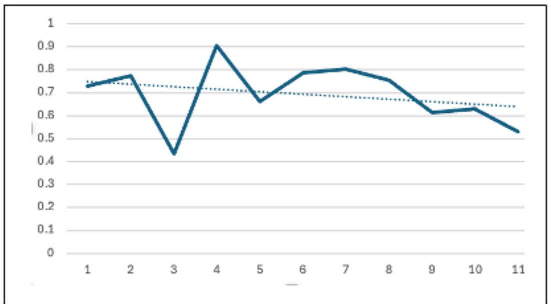


Figure 8

The three tables from figure 8 represent the outputs of RPISeq using an input of the subgroup of lncRNA sequences that were identified in the eQTL maize data as regulators of target genes with the states that are included in TG1 (GRMZM2G108265). With the states being

S10, S8, S7, S6, S4, S2, the subgroup of lncRNAs totaled to 11. The first table is the result of running the 11 lncRNA sequences in RPISeq against the protein sequence of the maize version of FKBP4 yielding an SVM and RF probability score for each lncRNA. The second table is the result of running the 11 lncRNA sequences in RPISeq against the protein sequence of the maize version of TAF15 yielding an SVM and RF probability score for each lncRNA. The third table is the result of running the 11 lncRNA sequences in RPISeq against the protein sequence of the maize version of WDR43 yielding an SVM and RF probability score for each lncRNA. The green highlights represent the highest values in each classifier and the red highlights represent the lowest values.

The graphs next to each table visually represent the average SVM and RF score for binding with each protein in the L1 subgroup. These average SVM and RF scores are also represented in a table (Figure 10).

a.

In the first table, GRMZM2G041842 had the highest SVM (.967) and RF (.8) values, meaning it was the lncRNA predicted to be the most likely to bind to FKBP4. The least likely to bind to FKBP4 was predicted to be GRMZM2G027825, because it had the lowest SVM (.457) and RF (.55) values.

b.

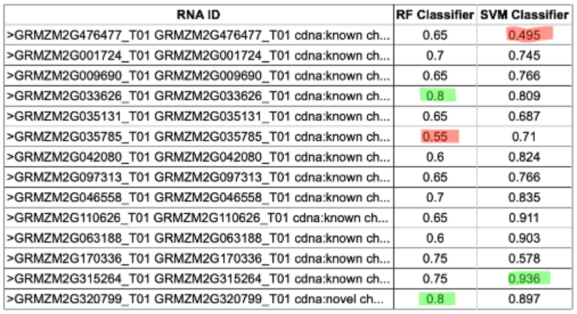
In the second table, GRMZM2G041842 had the highest SVM (.959), meaning it was the lncRNA predicted to be the most likely to bind to TAF15 according to the SVM classifier. According to the RF classifier, GRMZM2G018006, GRMZM2G041842, and GRMZM2G076636 were the lncRNAs predicted to be the most likely to bind to TAF15 since they all had RF values of .85. The least likely to bind to TAF15 according to the SVM classifier was predicted to be GRZM2G374313, because it had the lowest SVM, with a value of .466. GRMZM2G027825 and GRMZM2G374313 had values of .6 according to the RF classifier, which was the lowest values out of the other lncRNAs, making them predicted to be the least likely to bind to TAF15.

c.

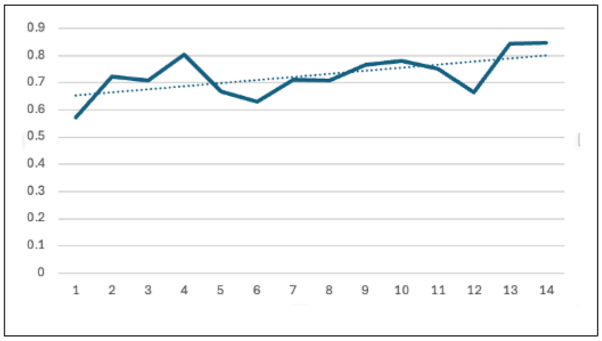
In the third table, GRMZM2G041842 had the highest SVM value (.961), meaning it was the lncRNA most likely to bind to WDR43 according to the SVM classifier. GRMZM2G076636 had the highest RF value (.9), meaning it was the lncRNA most likely to bind to WDR43 according to the RF classifier. GRMZM2G027825 was the lncRNA predicted to be the least likely to bind to WDR43 since it had the lowest SVM (.368) and RF (.5) values.

## RPISeq Predictions for L2

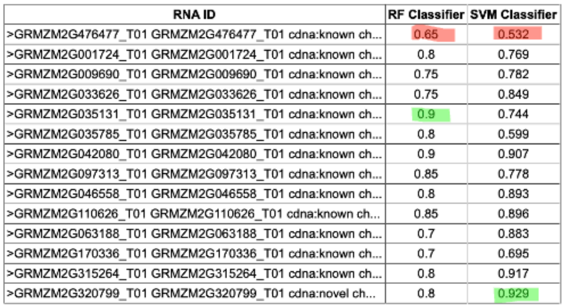
a. FKBP4/L2



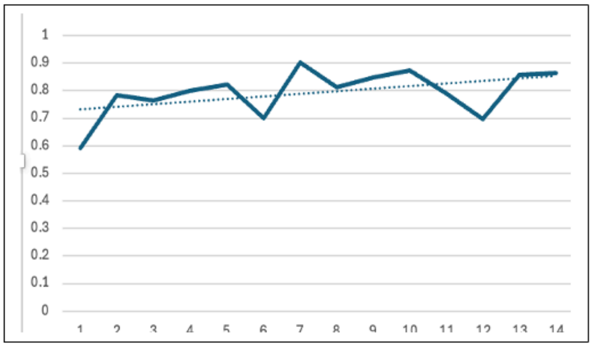
| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G476477_T01 GRMZM2G476477_T01 cdna:known ch... | 0.65 | 0.495 |
| >GRMZM2G001724_T01 GRMZM2G001724_T01 cdna:known ch... | 0.7 | 0.745 |
| >GRMZM2G009690_T01 GRMZM2G009690_T01 cdna:known ch... | 0.65 | 0.766 |
| >GRMZM2G033626_T01 GRMZM2G033626_T01 cdna:known ch... | 0.8 | 0.809 |
| >GRMZM2G035131_T01 GRMZM2G035131_T01 cdna:known ch... | 0.65 | 0.687 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.55 | 0.71 |
| >GRMZM2G042080_T01 GRMZM2G042080_T01 cdna:known ch... | 0.6 | 0.824 |
| >GRMZM2G097313_T01 GRMZM2G097313_T01 cdna:known ch... | 0.65 | 0.766 |
| >GRMZM2G046558_T01 GRMZM2G046558_T01 cdna:known ch... | 0.7 | 0.835 |
| >GRMZM2G110626_T01 GRMZM2G110626_T01 cdna:known ch... | 0.65 | 0.911 |
| >GRMZM2G063188_T01 GRMZM2G063188_T01 cdna:known ch... | 0.6 | 0.903 |
| >GRMZM2G170336_T01 GRMZM2G170336_T01 cdna:known ch... | 0.75 | 0.578 |
| >GRMZM2G315264_T01 GRMZM2G315264_T01 cdna:known ch... | 0.75 | 0.936 |
| >GRMZM2G320799_T01 GRMZM2G320799_T01 cdna:novel ch... | 0.8 | 0.897 |



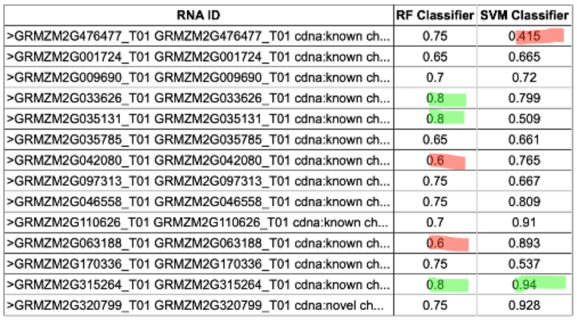
b. TAF15/L2



| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G476477_T01 GRMZM2G476477_T01 cdna:known ch... | 0.65 | 0.532 |
| >GRMZM2G001724_T01 GRMZM2G001724_T01 cdna:known ch... | 0.8 | 0.769 |
| >GRMZM2G009690_T01 GRMZM2G009690_T01 cdna:known ch... | 0.75 | 0.782 |
| >GRMZM2G033626_T01 GRMZM2G033626_T01 cdna:known ch... | 0.75 | 0.849 |
| >GRMZM2G035131_T01 GRMZM2G035131_T01 cdna:known ch... | 0.9 | 0.744 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.8 | 0.599 |
| >GRMZM2G042080_T01 GRMZM2G042080_T01 cdna:known ch... | 0.9 | 0.907 |
| >GRMZM2G097313_T01 GRMZM2G097313_T01 cdna:known ch... | 0.85 | 0.778 |
| >GRMZM2G046558_T01 GRMZM2G046558_T01 cdna:known ch... | 0.8 | 0.893 |
| >GRMZM2G110626_T01 GRMZM2G110626_T01 cdna:known ch... | 0.85 | 0.896 |
| >GRMZM2G063188_T01 GRMZM2G063188_T01 cdna:known ch... | 0.7 | 0.883 |
| >GRMZM2G170336_T01 GRMZM2G170336_T01 cdna:known ch... | 0.7 | 0.695 |
| >GRMZM2G315264_T01 GRMZM2G315264_T01 cdna:known ch... | 0.8 | 0.917 |
| >GRMZM2G320799_T01 GRMZM2G320799_T01 cdna:novel ch... | 0.8 | 0.929 |



c. WDR43/L2



| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G476477_T01 GRMZM2G476477_T01 cdna:known ch... | 0.75 | 0.415 |
| >GRMZM2G001724_T01 GRMZM2G001724_T01 cdna:known ch... | 0.65 | 0.665 |
| >GRMZM2G009690_T01 GRMZM2G009690_T01 cdna:known ch... | 0.7 | 0.72 |
| >GRMZM2G033626_T01 GRMZM2G033626_T01 cdna:known ch... | 0.8 | 0.799 |
| >GRMZM2G035131_T01 GRMZM2G035131_T01 cdna:known ch... | 0.8 | 0.509 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.65 | 0.661 |
| >GRMZM2G042080_T01 GRMZM2G042080_T01 cdna:known ch... | 0.6 | 0.765 |
| >GRMZM2G097313_T01 GRMZM2G097313_T01 cdna:known ch... | 0.75 | 0.667 |
| >GRMZM2G046558_T01 GRMZM2G046558_T01 cdna:known ch... | 0.75 | 0.809 |
| >GRMZM2G110626_T01 GRMZM2G110626_T01 cdna:known ch... | 0.7 | 0.91 |
| >GRMZM2G063188_T01 GRMZM2G063188_T01 cdna:known ch... | 0.6 | 0.893 |
| >GRMZM2G170336_T01 GRMZM2G170336_T01 cdna:known ch... | 0.75 | 0.537 |
| >GRMZM2G315264_T01 GRMZM2G315264_T01 cdna:known ch... | 0.8 | 0.94 |
| >GRMZM2G320799_T01 GRMZM2G320799_T01 cdna:novel ch... | 0.75 | 0.928 |

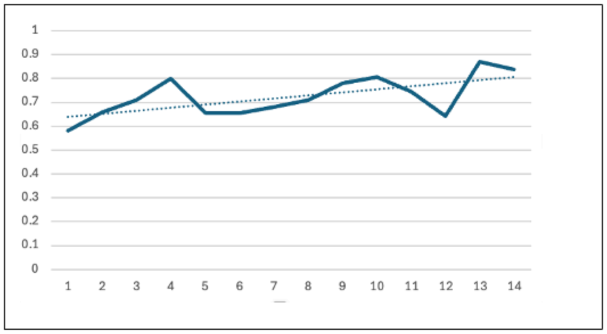


Figure 9

The three tables from figure 9 represent the outputs of RPISeq using an input of the subgroup of lncRNA sequences that were identified in the eQTL maize data as regulators of target genes with the states that are included in TG2 (GRMZM2G041694). With the states being S25, S17, S1, S5, S4, the subgroup of lncRNAs totaled to 14. The first table is the result of running the 14 lncRNA sequences in RPISeq against the protein sequence of the maize version of FKBP4 yielding an SVM and RF probability score for each lncRNA. The second table is the result of running the 14 lncRNA sequences in RPISeq against the protein sequence of the maize version of TAF15 yielding an SVM and RF probability score for each lncRNA. The third table is the result of running the 14 lncRNA sequences in RPISeq against the protein sequence of the maize version of WDR43 yielding an SVM and RF probability score for each lncRNA. The green highlights represent the highest values in each classifier and the red highlights represent the lowest values.

The graphs next to each table visually represent the average SVM and RF score for binding with each protein in the L2 subgroup. These average SVM and RF scores are also represented in a table (Figure 10).

a.

In the first table, GRMZM2G315264 had the highest SVM (.936), meaning it was the lncRNA predicted to be the most likely to bind to FKBP4 according to the SVM classifier. The most likely to bind to FKBP4 according to the RF classifier were GRMZM2G033626 and GRMZM2G320799, which had RF values of .8. GRMZM2G476477 was predicted to be the least likely to bind with FKBP4 since it had the lowest SVM (.495). The lncRNA with the lowest RF value (.55) was GRMZM2G035785.

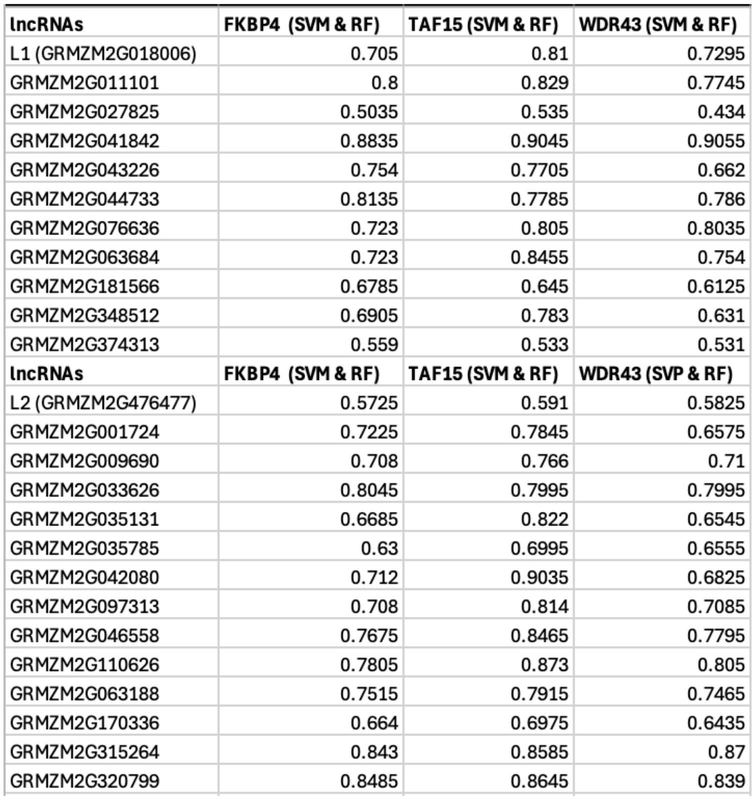
b.

In the second table, GRMZM2G320799 had the highest SVM (.929), meaning it was the lncRNA predicted to be the most likely to bind to TAF15 according to the SVM classifier. According to the RF classifier, GRMZM2G035131 was the lncRNA predicted to be the most likely to bind to TAF15 since it had an RF value of .9. The least likely to bind to TAF15 was predicted to be GRMZM2G476477 since it had an SVM value of .532 and RF value of .65, which were the lowest values for each category.

c.

In the third table, GRMZM2G320799 had the highest SVM value (.94), meaning it was the lncRNA most likely to bind to WDR43 according to the SVM classifier. GRMZM2G033626, GRMZM2G035131, and GRMZM2G315264 had the highest RF values (.8), meaning they were the lncRNAs most likely to bind to WDR43 according to the RF classifier. GRMZM2G476477 had the lowest SVM values so it was the least likely to bind to WDR43 according to the SVM classifier. GRMZM2G042080 and GRMZM2G064188 had RF values of .6, making them predicted to be the least likely lncRNAs to bind to WDR43 according to the RF classifier.

**Average lncRNA Predictions from RPISeq**



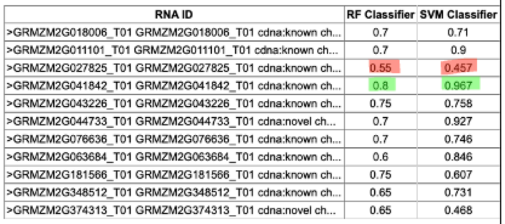
| lncRNAs | FKBP4 (SVM & RF) | TAF15 (SVM & RF) | WDR43 (SVM & RF) |
|---|---|---|---|
| L1 (GRMZM2G018006) | 0.705 | 0.81 | 0.7295 |
| GRMZM2G011101 | 0.8 | 0.829 | 0.7745 |
| GRMZM2G027825 | 0.5035 | 0.535 | 0.434 |
| GRMZM2G041842 | 0.8835 | 0.9045 | 0.9055 |
| GRMZM2G043226 | 0.754 | 0.7705 | 0.662 |
| GRMZM2G044733 | 0.8135 | 0.7785 | 0.786 |
| GRMZM2G076636 | 0.723 | 0.805 | 0.8035 |
| GRMZM2G063684 | 0.723 | 0.8455 | 0.754 |
| GRMZM2G181566 | 0.6785 | 0.645 | 0.6125 |
| GRMZM2G348512 | 0.6905 | 0.783 | 0.631 |
| GRMZM2G374313 | 0.559 | 0.533 | 0.531 |
| **lncRNAs** | **FKBP4 (SVM & RF)** | **TAF15 (SVM & RF)** | **WDR43 (SVP & RF)** |
| L2 (GRMZM2G476477) | 0.5725 | 0.591 | 0.5825 |
| GRMZM2G001724 | 0.7225 | 0.7845 | 0.6575 |
| GRMZM2G009690 | 0.708 | 0.766 | 0.71 |
| GRMZM2G033626 | 0.8045 | 0.7995 | 0.7995 |
| GRMZM2G035131 | 0.6685 | 0.822 | 0.6545 |
| GRMZM2G035785 | 0.63 | 0.6995 | 0.6555 |
| GRMZM2G042080 | 0.712 | 0.9035 | 0.6825 |
| GRMZM2G097313 | 0.708 | 0.814 | 0.7085 |
| GRMZM2G046558 | 0.7675 | 0.8465 | 0.7795 |
| GRMZM2G110626 | 0.7805 | 0.873 | 0.805 |
| GRMZM2G063188 | 0.7515 | 0.7915 | 0.7465 |
| GRMZM2G170336 | 0.664 | 0.6975 | 0.6435 |
| GRMZM2G315264 | 0.843 | 0.8585 | 0.87 |
| GRMZM2G320799 | 0.8485 | 0.8645 | 0.839 |

Figure 10

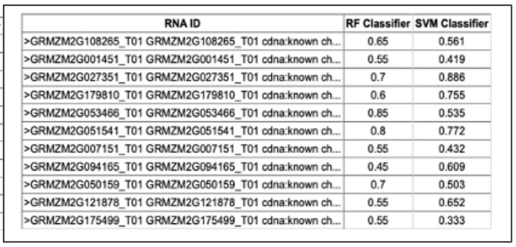
The above figure is a table that displays the gene IDs for all the lncRNAs implemented into RPISeq and the combined average of SVM and RF values for each lncRNA. The first 12 lncRNAs are those that were identified as regulators from the maize eQTL data that regulate the target genes that that have the same chromatin states as TG1 and the last 18 are those that were identified as regulators from the maize eQTL data that regulate the target genes that that have the same chromatin states as TG2. Each value is the combined average of SVM and RF for each protein in each lncRNA.

## RPISeq Predictions for L1 and TG1

a. FKBP4/L1/TG1

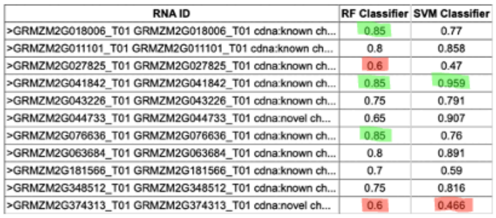


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G018006_T01 GRMZM2G018006_T01 cdna:known ch... | 0.7 | 0.71 |
| >GRMZM2G011101_T01 GRMZM2G011101_T01 cdna:known ch... | 0.7 | 0.9 |
| >GRMZM2G027825_T01 GRMZM2G027825_T01 cdna:known ch... | 0.55 | 0.457 |
| >GRMZM2G041842_T01 GRMZM2G041842_T01 cdna:known ch... | 0.8 | 0.967 |
| >GRMZM2G043226_T01 GRMZM2G043226_T01 cdna:known ch... | 0.75 | 0.758 |
| >GRMZM2G044733_T01 GRMZM2G044733_T01 cdna:novel ch... | 0.7 | 0.927 |
| >GRMZM2G076636_T01 GRMZM2G076636_T01 cdna:known ch... | 0.7 | 0.746 |
| >GRMZM2G063684_T01 GRMZM2G063684_T01 cdna:known ch... | 0.6 | 0.846 |
| >GRMZM2G181566_T01 GRMZM2G181566_T01 cdna:known ch... | 0.75 | 0.607 |
| >GRMZM2G348512_T01 GRMZM2G348512_T01 cdna:known ch... | 0.65 | 0.731 |
| >GRMZM2G374313_T01 GRMZM2G374313_T01 cdna:novel ch... | 0.65 | 0.468 |

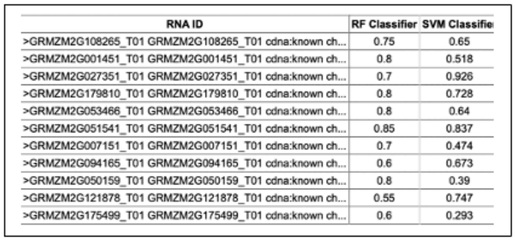


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G108265_T01 GRMZM2G108265_T01 cdna:known ch... | 0.65 | 0.561 |
| >GRMZM2G001451_T01 GRMZM2G001451_T01 cdna:known ch... | 0.55 | 0.419 |
| >GRMZM2G027351_T01 GRMZM2G027351_T01 cdna:known ch... | 0.7 | 0.886 |
| >GRMZM2G179810_T01 GRMZM2G179810_T01 cdna:known ch... | 0.6 | 0.755 |
| >GRMZM2G053466_T01 GRMZM2G053466_T01 cdna:known ch... | 0.85 | 0.535 |
| >GRMZM2G051541_T01 GRMZM2G051541_T01 cdna:known ch... | 0.8 | 0.772 |
| >GRMZM2G007151_T01 GRMZM2G007151_T01 cdna:known ch... | 0.55 | 0.432 |
| >GRMZM2G094165_T01 GRMZM2G094165_T01 cdna:known ch... | 0.45 | 0.609 |
| >GRMZM2G050159_T01 GRMZM2G050159_T01 cdna:known ch... | 0.7 | 0.503 |
| >GRMZM2G121878_T01 GRMZM2G121878_T01 cdna:known ch... | 0.55 | 0.652 |
| >GRMZM2G175499_T01 GRMZM2G175499_T01 cdna:known ch... | 0.55 | 0.333 |

b. TAF15/L1/TG1

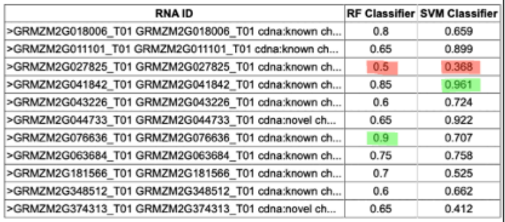


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G018006_T01 GRMZM2G018006_T01 cdna:known ch... | 0.85 | 0.77 |
| >GRMZM2G011101_T01 GRMZM2G011101_T01 cdna:known ch... | 0.8 | 0.858 |
| >GRMZM2G027825_T01 GRMZM2G027825_T01 cdna:known ch... | 0.6 | 0.47 |
| >GRMZM2G041842_T01 GRMZM2G041842_T01 cdna:known ch... | 0.85 | 0.959 |
| >GRMZM2G043226_T01 GRMZM2G043226_T01 cdna:known ch... | 0.75 | 0.791 |
| >GRMZM2G044733_T01 GRMZM2G044733_T01 cdna:novel ch... | 0.65 | 0.907 |
| >GRMZM2G076636_T01 GRMZM2G076636_T01 cdna:known ch... | 0.85 | 0.76 |
| >GRMZM2G063684_T01 GRMZM2G063684_T01 cdna:known ch... | 0.8 | 0.891 |
| >GRMZM2G181566_T01 GRMZM2G181566_T01 cdna:known ch... | 0.7 | 0.59 |
| >GRMZM2G348512_T01 GRMZM2G348512_T01 cdna:known ch... | 0.75 | 0.816 |
| >GRMZM2G374313_T01 GRMZM2G374313_T01 cdna:novel ch... | 0.6 | 0.466 |

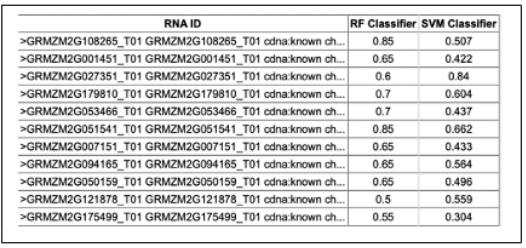


| RNA ID | RF Classifier | SVM Classifie |
|---|---|---|
| >GRMZM2G108265_T01 GRMZM2G108265_T01 cdna:known ch... | 0.75 | 0.65 |
| >GRMZM2G001451_T01 GRMZM2G001451_T01 cdna:known ch... | 0.8 | 0.518 |
| >GRMZM2G027351_T01 GRMZM2G027351_T01 cdna:known ch... | 0.7 | 0.926 |
| >GRMZM2G179810_T01 GRMZM2G179810_T01 cdna:known ch... | 0.8 | 0.728 |
| >GRMZM2G053466_T01 GRMZM2G053466_T01 cdna:known ch... | 0.8 | 0.64 |
| >GRMZM2G051541_T01 GRMZM2G051541_T01 cdna:known ch... | 0.85 | 0.837 |
| >GRMZM2G007151_T01 GRMZM2G007151_T01 cdna:known ch... | 0.7 | 0.474 |
| >GRMZM2G094165_T01 GRMZM2G094165_T01 cdna:known ch... | 0.6 | 0.673 |
| >GRMZM2G050159_T01 GRMZM2G050159_T01 cdna:known ch... | 0.8 | 0.39 |
| >GRMZM2G121878_T01 GRMZM2G121878_T01 cdna:known ch... | 0.55 | 0.747 |
| >GRMZM2G175499_T01 GRMZM2G175499_T01 cdna:known ch... | 0.6 | 0.293 |

c. WDR43/L1/TG1



| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G018006_T01 GRMZM2G018006_T01 cdna:known ch... | 0.8 | 0.659 |
| >GRMZM2G011101_T01 GRMZM2G011101_T01 cdna:known ch... | 0.65 | 0.899 |
| >GRMZM2G027825_T01 GRMZM2G027825_T01 cdna:known ch... | 0.5 | 0.368 |
| >GRMZM2G041842_T01 GRMZM2G041842_T01 cdna:known ch... | 0.85 | 0.961 |
| >GRMZM2G043226_T01 GRMZM2G043226_T01 cdna:known ch... | 0.6 | 0.724 |
| >GRMZM2G044733_T01 GRMZM2G044733_T01 cdna:novel ch... | 0.65 | 0.922 |
| >GRMZM2G076636_T01 GRMZM2G076636_T01 cdna:known ch... | 0.9 | 0.707 |
| >GRMZM2G063684_T01 GRMZM2G063684_T01 cdna:known ch... | 0.75 | 0.758 |
| >GRMZM2G181566_T01 GRMZM2G181566_T01 cdna:known ch... | 0.7 | 0.525 |
| >GRMZM2G348512_T01 GRMZM2G348512_T01 cdna:known ch... | 0.6 | 0.662 |
| >GRMZM2G374313_T01 GRMZM2G374313_T01 cdna:novel ch... | 0.65 | 0.412 |



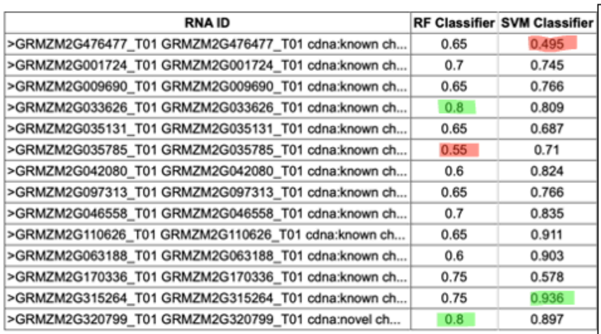
| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G108265_T01 GRMZM2G108265_T01 cdna:known ch... | 0.85 | 0.507 |
| >GRMZM2G001451_T01 GRMZM2G001451_T01 cdna:known ch... | 0.65 | 0.422 |
| >GRMZM2G027351_T01 GRMZM2G027351_T01 cdna:known ch... | 0.6 | 0.84 |
| >GRMZM2G179810_T01 GRMZM2G179810_T01 cdna:known ch... | 0.7 | 0.604 |
| >GRMZM2G053466_T01 GRMZM2G053466_T01 cdna:known ch... | 0.7 | 0.437 |
| >GRMZM2G051541_T01 GRMZM2G051541_T01 cdna:known ch... | 0.85 | 0.662 |
| >GRMZM2G007151_T01 GRMZM2G007151_T01 cdna:known ch... | 0.65 | 0.433 |
| >GRMZM2G094165_T01 GRMZM2G094165_T01 cdna:known ch... | 0.65 | 0.564 |
| >GRMZM2G050159_T01 GRMZM2G050159_T01 cdna:known ch... | 0.65 | 0.496 |
| >GRMZM2G121878_T01 GRMZM2G121878_T01 cdna:known ch... | 0.5 | 0.559 |
| >GRMZM2G175499_T01 GRMZM2G175499_T01 cdna:known ch... | 0.55 | 0.304 |

Figure 11

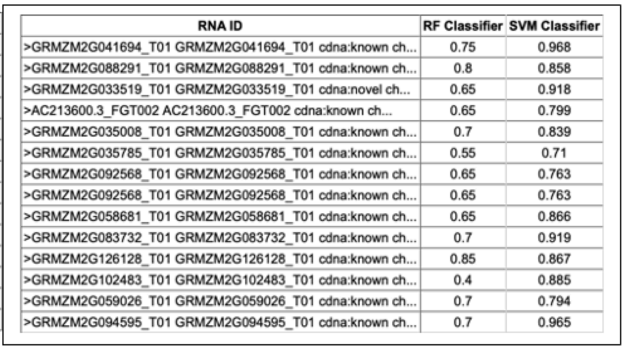
This figure above shows a side-by-side comparison of running L1 and TG1 subgroups in RPISeq. The tables to the left are the results from running L1, which are also seen in (Figure 8). The tables to the right are the results from running TG1.

## RPISeq Predictions for L2 and TG2

### a. FKBP4/L2/TG2

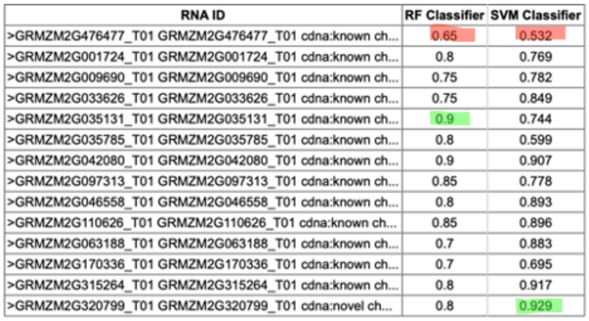


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G476477_T01 GRMZM2G476477_T01 cdna:known ch... | 0.65 | 0.495 |
| >GRMZM2G001724_T01 GRMZM2G001724_T01 cdna:known ch... | 0.7 | 0.745 |
| >GRMZM2G009690_T01 GRMZM2G009690_T01 cdna:known ch... | 0.65 | 0.766 |
| >GRMZM2G033626_T01 GRMZM2G033626_T01 cdna:known ch... | 0.8 | 0.809 |
| >GRMZM2G035131_T01 GRMZM2G035131_T01 cdna:known ch... | 0.65 | 0.687 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.55 | 0.71 |
| >GRMZM2G042080_T01 GRMZM2G042080_T01 cdna:known ch... | 0.6 | 0.824 |
| >GRMZM2G097313_T01 GRMZM2G097313_T01 cdna:known ch... | 0.65 | 0.766 |
| >GRMZM2G046558_T01 GRMZM2G046558_T01 cdna:known ch... | 0.7 | 0.835 |
| >GRMZM2G110626_T01 GRMZM2G110626_T01 cdna:known ch... | 0.65 | 0.911 |
| >GRMZM2G063188_T01 GRMZM2G063188_T01 cdna:known ch... | 0.6 | 0.903 |
| >GRMZM2G170336_T01 GRMZM2G170336_T01 cdna:known ch... | 0.75 | 0.578 |
| >GRMZM2G315264_T01 GRMZM2G315264_T01 cdna:known ch... | 0.75 | 0.936 |
| >GRMZM2G320799_T01 GRMZM2G320799_T01 cdna:novel ch... | 0.8 | 0.897 |

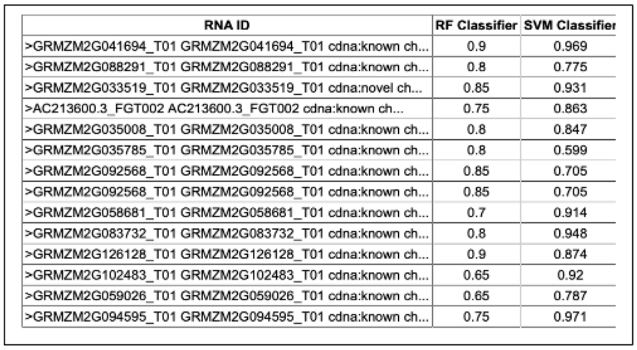


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G041694_T01 GRMZM2G041694_T01 cdna:known ch... | 0.75 | 0.968 |
| >GRMZM2G088291_T01 GRMZM2G088291_T01 cdna:known ch... | 0.8 | 0.858 |
| >GRMZM2G033519_T01 GRMZM2G033519_T01 cdna:novel ch... | 0.65 | 0.918 |
| >AC213600.3_FGT002 AC213600.3_FGT002 cdna:known ch... | 0.65 | 0.799 |
| >GRMZM2G035008_T01 GRMZM2G035008_T01 cdna:known ch... | 0.7 | 0.839 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.55 | 0.71 |
| >GRMZM2G092568_T01 GRMZM2G092568_T01 cdna:known ch... | 0.65 | 0.763 |
| >GRMZM2G092568_T01 GRMZM2G092568_T01 cdna:known ch... | 0.65 | 0.763 |
| >GRMZM2G058681_T01 GRMZM2G058681_T01 cdna:known ch... | 0.65 | 0.866 |
| >GRMZM2G083732_T01 GRMZM2G083732_T01 cdna:known ch... | 0.7 | 0.919 |
| >GRMZM2G126128_T01 GRMZM2G126128_T01 cdna:known ch... | 0.85 | 0.867 |
| >GRMZM2G102483_T01 GRMZM2G102483_T01 cdna:known ch... | 0.4 | 0.885 |
| >GRMZM2G059026_T01 GRMZM2G059026_T01 cdna:known ch... | 0.7 | 0.794 |
| >GRMZM2G094595_T01 GRMZM2G094595_T01 cdna:known ch... | 0.7 | 0.965 |

### b. TAF15/L2/TG2

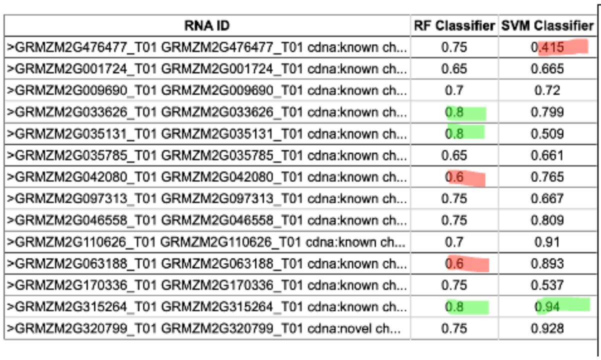


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G476477_T01 GRMZM2G476477_T01 cdna:known ch... | 0.65 | 0.532 |
| >GRMZM2G001724_T01 GRMZM2G001724_T01 cdna:known ch... | 0.8 | 0.769 |
| >GRMZM2G009690_T01 GRMZM2G009690_T01 cdna:known ch... | 0.75 | 0.782 |
| >GRMZM2G033626_T01 GRMZM2G033626_T01 cdna:known ch... | 0.75 | 0.849 |
| >GRMZM2G035131_T01 GRMZM2G035131_T01 cdna:known ch... | 0.9 | 0.744 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.8 | 0.599 |
| >GRMZM2G042080_T01 GRMZM2G042080_T01 cdna:known ch... | 0.9 | 0.907 |
| >GRMZM2G097313_T01 GRMZM2G097313_T01 cdna:known ch... | 0.85 | 0.778 |
| >GRMZM2G046558_T01 GRMZM2G046558_T01 cdna:known ch... | 0.8 | 0.893 |
| >GRMZM2G110626_T01 GRMZM2G110626_T01 cdna:known ch... | 0.85 | 0.896 |
| >GRMZM2G063188_T01 GRMZM2G063188_T01 cdna:known ch... | 0.7 | 0.883 |
| >GRMZM2G170336_T01 GRMZM2G170336_T01 cdna:known ch... | 0.7 | 0.695 |
| >GRMZM2G315264_T01 GRMZM2G315264_T01 cdna:known ch... | 0.8 | 0.917 |
| >GRMZM2G320799_T01 GRMZM2G320799_T01 cdna:novel ch... | 0.8 | 0.929 |

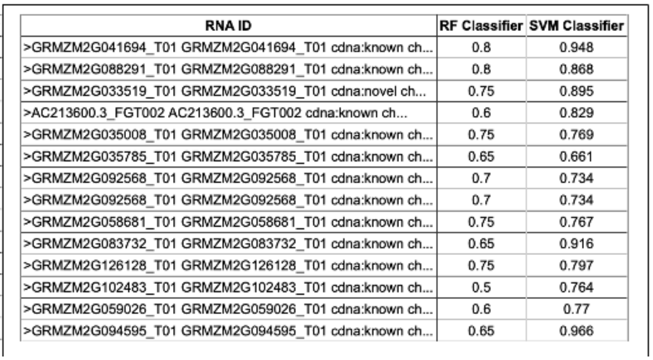


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G041694_T01 GRMZM2G041694_T01 cdna:known ch... | 0.9 | 0.969 |
| >GRMZM2G088291_T01 GRMZM2G088291_T01 cdna:known ch... | 0.8 | 0.775 |
| >GRMZM2G033519_T01 GRMZM2G033519_T01 cdna:novel ch... | 0.85 | 0.931 |
| >AC213600.3_FGT002 AC213600.3_FGT002 cdna:known ch... | 0.75 | 0.863 |
| >GRMZM2G035008_T01 GRMZM2G035008_T01 cdna:known ch... | 0.8 | 0.847 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.8 | 0.599 |
| >GRMZM2G092568_T01 GRMZM2G092568_T01 cdna:known ch... | 0.85 | 0.705 |
| >GRMZM2G092568_T01 GRMZM2G092568_T01 cdna:known ch... | 0.85 | 0.705 |
| >GRMZM2G058681_T01 GRMZM2G058681_T01 cdna:known ch... | 0.7 | 0.914 |
| >GRMZM2G083732_T01 GRMZM2G083732_T01 cdna:known ch... | 0.8 | 0.948 |
| >GRMZM2G126128_T01 GRMZM2G126128_T01 cdna:known ch... | 0.9 | 0.874 |
| >GRMZM2G102483_T01 GRMZM2G102483_T01 cdna:known ch... | 0.65 | 0.92 |
| >GRMZM2G059026_T01 GRMZM2G059026_T01 cdna:known ch... | 0.65 | 0.787 |
| >GRMZM2G094595_T01 GRMZM2G094595_T01 cdna:known ch... | 0.75 | 0.971 |

### c. WDR43/L2/TG2



| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G476477_T01 GRMZM2G476477_T01 cdna:known ch... | 0.75 | 0.415 |
| >GRMZM2G001724_T01 GRMZM2G001724_T01 cdna:known ch... | 0.65 | 0.665 |
| >GRMZM2G009690_T01 GRMZM2G009690_T01 cdna:known ch... | 0.7 | 0.72 |
| >GRMZM2G033626_T01 GRMZM2G033626_T01 cdna:known ch... | 0.8 | 0.799 |
| >GRMZM2G035131_T01 GRMZM2G035131_T01 cdna:known ch... | 0.8 | 0.509 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.65 | 0.661 |
| >GRMZM2G042080_T01 GRMZM2G042080_T01 cdna:known ch... | 0.6 | 0.765 |
| >GRMZM2G097313_T01 GRMZM2G097313_T01 cdna:known ch... | 0.75 | 0.667 |
| >GRMZM2G046558_T01 GRMZM2G046558_T01 cdna:known ch... | 0.75 | 0.809 |
| >GRMZM2G110626_T01 GRMZM2G110626_T01 cdna:known ch... | 0.7 | 0.91 |
| >GRMZM2G063188_T01 GRMZM2G063188_T01 cdna:known ch... | 0.6 | 0.893 |
| >GRMZM2G170336_T01 GRMZM2G170336_T01 cdna:known ch... | 0.75 | 0.537 |
| >GRMZM2G315264_T01 GRMZM2G315264_T01 cdna:known ch... | 0.8 | 0.94 |
| >GRMZM2G320799_T01 GRMZM2G320799_T01 cdna:novel ch... | 0.75 | 0.928 |

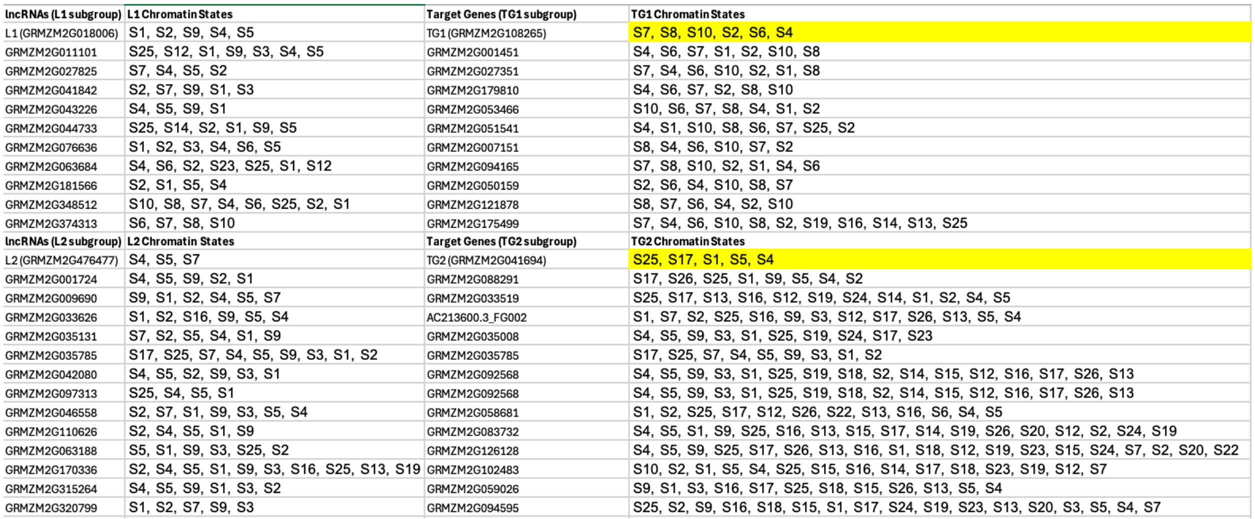


| RNA ID | RF Classifier | SVM Classifier |
|---|---|---|
| >GRMZM2G041694_T01 GRMZM2G041694_T01 cdna:known ch... | 0.8 | 0.948 |
| >GRMZM2G088291_T01 GRMZM2G088291_T01 cdna:known ch... | 0.8 | 0.868 |
| >GRMZM2G033519_T01 GRMZM2G033519_T01 cdna:novel ch... | 0.75 | 0.895 |
| >AC213600.3_FGT002 AC213600.3_FGT002 cdna:known ch... | 0.6 | 0.829 |
| >GRMZM2G035008_T01 GRMZM2G035008_T01 cdna:known ch... | 0.75 | 0.769 |
| >GRMZM2G035785_T01 GRMZM2G035785_T01 cdna:known ch... | 0.65 | 0.661 |
| >GRMZM2G092568_T01 GRMZM2G092568_T01 cdna:known ch... | 0.7 | 0.734 |
| >GRMZM2G092568_T01 GRMZM2G092568_T01 cdna:known ch... | 0.7 | 0.734 |
| >GRMZM2G058681_T01 GRMZM2G058681_T01 cdna:known ch... | 0.75 | 0.767 |
| >GRMZM2G083732_T01 GRMZM2G083732_T01 cdna:known ch... | 0.65 | 0.916 |
| >GRMZM2G126128_T01 GRMZM2G126128_T01 cdna:known ch... | 0.75 | 0.797 |
| >GRMZM2G102483_T01 GRMZM2G102483_T01 cdna:known ch... | 0.5 | 0.764 |
| >GRMZM2G059026_T01 GRMZM2G059026_T01 cdna:known ch... | 0.6 | 0.77 |
| >GRMZM2G094595_T01 GRMZM2G094595_T01 cdna:known ch... | 0.65 | 0.966 |

Figure 12

This figure above shows a side-by-side comparison of running L2 and TG2 subgroups in RPISeq. The tables to the left are the results from running L2, which are also seen in (Figure 9). The tables to the right are the results from running TG2.

## Chromatin States of L and TG Subgroups

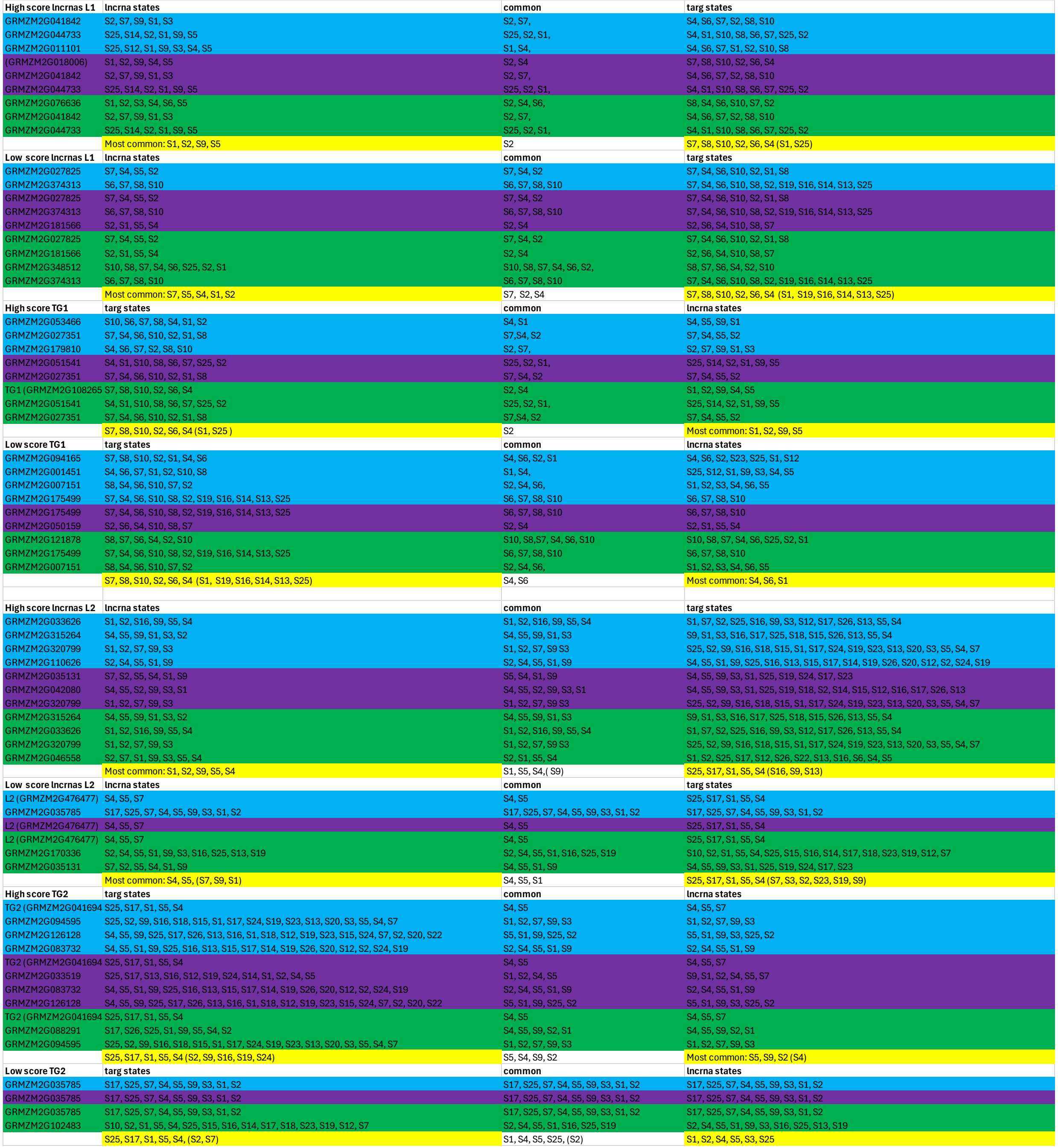


| lncRNAs (L1 subgroup) | L1 Chromatin States | Target Genes (TG1 subgroup) | TG1 Chromatin States |
|---|---|---|---|
| L1 (GRMZM2G018006) | S1, S2, S9, S4, S5 | TG1 (GRMZM2G108265) | S7, S8, S10, S2, S6, S4 |
| GRMZM2G011101 | S25, S12, S1, S9, S3, S4, S5 | GRMZM2G001451 | S4, S6, S7, S1, S2, S10, S8 |
| GRMZM2G027825 | S7, S4, S5, S2 | GRMZM2G027351 | S7, S4, S6, S10, S2, S1, S8 |
| GRMZM2G041842 | S2, S7, S9, S1, S3 | GRMZM2G179810 | S4, S6, S7, S2, S8, S10 |
| GRMZM2G043226 | S4, S5, S9, S1 | GRMZM2G053466 | S10, S6, S7, S8, S4, S1, S2 |
| GRMZM2G044733 | S25, S14, S2, S1, S9, S5 | GRMZM2G051541 | S4, S1, S10, S8, S6, S7, S25, S2 |
| GRMZM2G076636 | S1, S2, S3, S4, S6, S5 | GRMZM2G007151 | S8, S4, S6, S10, S7, S2 |
| GRMZM2G063684 | S4, S6, S2, S23, S25, S1, S12 | GRMZM2G094165 | S7, S8, S10, S2, S1, S4, S6 |
| GRMZM2G181566 | S2, S1, S5, S4 | GRMZM2G050159 | S2, S6, S4, S10, S8, S7 |
| GRMZM2G348512 | S10, S8, S7, S4, S6, S25, S2, S1 | GRMZM2G121878 | S8, S7, S6, S4, S2, S10 |
| GRMZM2G374313 | S6, S7, S8, S10 | GRMZM2G175499 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 |
| **lncRNAs (L2 subgroup)** | **L2 Chromatin States** | **Target Genes (TG2 subgroup)** | **TG2 Chromatin States** |
| L2 (GRMZM2G476477) | S4, S5, S7 | TG2 (GRMZM2G041694) | S25, S17, S1, S5, S4 |
| GRMZM2G001724 | S4, S5, S9, S2, S1 | GRMZM2G088291 | S17, S26, S25, S1, S9, S5, S4, S2 |
| GRMZM2G009690 | S9, S1, S2, S4, S5, S7 | GRMZM2G033519 | S25, S17, S13, S16, S12, S19, S24, S14, S1, S2, S4, S5 |
| GRMZM2G033626 | S1, S2, S16, S9, S5, S4 | AC213600.3_FG002 | S1, S7, S2, S25, S16, S9, S3, S12, S17, S26, S13, S5, S4 |
| GRMZM2G035131 | S7, S2, S5, S4, S1, S9 | GRMZM2G035008 | S4, S5, S9, S3, S1, S25, S19, S24, S17, S23 |
| GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 |
| GRMZM2G042080 | S4, S5, S2, S9, S3, S1 | GRMZM2G092568 | S4, S5, S9, S3, S1, S25, S19, S18, S2, S14, S15, S12, S16, S17, S26, S13 |
| GRMZM2G097313 | S25, S4, S5, S1 | GRMZM2G092568 | S4, S5, S9, S3, S1, S25, S19, S18, S2, S14, S15, S12, S16, S17, S26, S13 |
| GRMZM2G046558 | S2, S7, S1, S9, S3, S5, S4 | GRMZM2G058681 | S1, S2, S25, S17, S12, S26, S22, S13, S16, S6, S4, S5 |
| GRMZM2G110626 | S2, S4, S5, S1, S9 | GRMZM2G083732 | S4, S5, S1, S9, S25, S16, S13, S15, S17, S14, S19, S26, S20, S12, S2, S24, S19 |
| GRMZM2G063188 | S5, S1, S9, S3, S25, S2 | GRMZM2G126128 | S4, S5, S9, S25, S17, S26, S13, S16, S1, S18, S12, S19, S23, S15, S24, S7, S2, S20, S22 |
| GRMZM2G170336 | S2, S4, S5, S1, S9, S3, S16, S25, S13, S19 | GRMZM2G102483 | S10, S2, S1, S5, S4, S25, S15, S16, S14, S17, S18, S23, S19, S12, S7 |
| GRMZM2G315264 | S4, S5, S9, S1, S3, S2 | GRMZM2G059026 | S9, S1, S3, S16, S17, S25, S18, S15, S26, S13, S5, S4 |
| GRMZM2G320799 | S1, S2, S7, S9, S3 | GRMZM2G094595 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 |

Figure 13

This figure shows a side-by-side comparison of the chromatin states that are present in the lncRNA subgroups and the target gene subgroups, indicating the difference in arrangements. The yellow highlights represent the state common states in the TG subgroups, that were used to identify those chosen for the subgroups.

## Chromatin State Comparison Analysis of L and TG Subgroups



| High score lncrnas L1 | lncrna states | common | targ states |
|---|---|---|---|
| GRMZM2G041842 | S2, S7, S9, S1, S3 | S2, S7, | S4, S6, S7, S2, S8, S10 |
| GRMZM2G044733 | S25, S14, S2, S1, S9, S5 | S25, S2, S1, | S4, S1, S10, S8, S6, S7, S25, S2 |
| GRMZM2G011101 | S25, S12, S1, S9, S3, S4, S5 | S1, S4, | S4, S6, S7, S1, S2, S10, S8 |
| (GRMZM2G018006) | S1, S2, S9, S4, S5 | S2, S4 | S7, S8, S10, S2, S6, S4 |
| GRMZM2G041842 | S2, S7, S9, S1, S3 | S2, S7, | S4, S6, S7, S2, S8, S10 |
| GRMZM2G044733 | S25, S14, S2, S1, S9, S5 | S25, S2, S1, | S4, S1, S10, S8, S6, S7, S25, S2 |
| GRMZM2G076636 | S1, S2, S3, S4, S6, S5 | S2, S4, S6, | S8, S4, S6, S10, S7, S2 |
| GRMZM2G041842 | S2, S7, S9, S1, S3 | S2, S7, | S4, S6, S7, S2, S8, S10 |
| GRMZM2G044733 | S25, S14, S2, S1, S9, S5 | S25, S2, S1, | S4, S1, S10, S8, S6, S7, S25, S2 |
|  | Most common: S1, S2, S9, S5 | S2 | S7, S8, S10, S2, S6, S4 (S1, S25) |
| **Low score lncrnas L1** | **lncrna states** | **common** | **targ states** |
| GRMZM2G027825 | S7, S4, S5, S2 | S7, S4, S2 | S7, S4, S6, S10, S2, S1, S8 |
| GRMZM2G374313 | S6, S7, S8, S10 | S6, S7, S8, S10 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 |
| GRMZM2G027825 | S7, S4, S5, S2 | S7, S4, S2 | S7, S4, S6, S10, S2, S1, S8 |
| GRMZM2G374313 | S6, S7, S8, S10 | S6, S7, S8, S10 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 |
| GRMZM2G181566 | S2, S1, S5, S4 | S2, S4 | S2, S6, S4, S10, S8, S7 |
| GRMZM2G027825 | S7, S4, S5, S2 | S7, S4, S2 | S7, S4, S6, S10, S2, S1, S8 |
| GRMZM2G181566 | S2, S1, S5, S4 | S2, S4 | S2, S6, S4, S10, S8, S7 |
| GRMZM2G348512 | S10, S8, S7, S4, S6, S25, S2, S1 | S10, S8, S7, S4, S6, S2, | S8, S7, S6, S4, S2, S10 |
| GRMZM2G374313 | S6, S7, S8, S10 | S6, S7, S8, S10 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 |
|  | Most common: S7, S5, S4, S1, S2 | S7, S2, S4 | S7, S8, S10, S2, S6, S4 (S1, S19, S16, S14, S13, S25) |
| **High score TG1** | **targ states** | **common** | **lncrna states** |
| GRMZM2G053466 | S10, S6, S7, S8, S4, S1, S2 | S4, S1 | S4, S5, S9, S1 |
| GRMZM2G027351 | S7, S4, S6, S10, S2, S1, S8 | S7,S4, S2 | S7, S4, S5, S2 |
| GRMZM2G179810 | S4, S6, S7, S2, S8, S10 | S2, S7, | S2, S7, S9, S1, S3 |
| GRMZM2G051541 | S4, S1, S10, S8, S6, S7, S25, S2 | S25, S2, S1, | S25, S14, S2, S1, S9, S5 |
| GRMZM2G027351 | S7, S4, S6, S10, S2, S1, S8 | S7, S4, S2 | S7, S4, S5, S2 |
| TG1 (GRMZM2G108265 | S7, S8, S10, S2, S6, S4 | S2, S4 | S1, S2, S9, S4, S5 |
| GRMZM2G051541 | S4, S1, S10, S8, S6, S7, S25, S2 | S25, S2, S1, | S25, S14, S2, S1, S9, S5 |
| GRMZM2G027351 | S7, S4, S6, S10, S2, S1, S8 | S7,S4, S2 | S7, S4, S5, S2 |
|  | S7, S8, S10, S2, S6, S4 (S1, S25 ) | S2 | Most common: S1, S2, S9, S5 |
| **Low score TG1** | **targ states** | **common** | **lncrna states** |
| GRMZM2G094165 | S7, S8, S10, S2, S1, S4, S6 | S4, S6, S2, S1 | S4, S6, S2, S23, S25, S1, S12 |
| GRMZM2G001451 | S4, S6, S7, S1, S2, S10, S8 | S1, S4, | S25, S12, S1, S9, S3, S4, S5 |
| GRMZM2G007151 | S8, S4, S6, S10, S7, S2 | S2, S4, S6, | S1, S2, S3, S4, S6, S5 |
| GRMZM2G175499 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 | S6, S7, S8, S10 | S6, S7, S8, S10 |
| GRMZM2G175499 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 | S6, S7, S8, S10 | S6, S7, S8, S10 |
| GRMZM2G050159 | S2, S6, S4, S10, S8, S7 | S2, S4 | S2, S1, S5, S4 |
| GRMZM2G121878 | S8, S7, S6, S4, S2, S10 | S10, S8,S7, S4, S6, S10 | S10, S8, S7, S4, S6, S25, S2, S1 |
| GRMZM2G175499 | S7, S4, S6, S10, S8, S2, S19, S16, S14, S13, S25 | S6, S7, S8, S10 | S6, S7, S8, S10 |
| GRMZM2G007151 | S8, S4, S6, S10, S7, S2 | S2, S4, S6, | S1, S2, S3, S4, S6, S5 |
|  | S7, S8, S10, S2, S6, S4 (S1, S19, S16, S14, S13, S25) | S4, S6 | Most common: S4, S6, S1 |
|  |  |  |  |
| **High score lncrnas L2** | **lncrna states** | **common** | **targ states** |
| GRMZM2G033626 | S1, S2, S16, S9, S5, S4 | S1, S2, S16, S9, S5, S4 | S1, S7, S2, S25, S16, S9, S3, S12, S17, S26, S13, S5, S4 |
| GRMZM2G315264 | S4, S5, S9, S1, S3, S2 | S4, S5, S9, S1, S3 | S9, S1, S3, S16, S17, S25, S18, S15, S26, S13, S5, S4 |
| GRMZM2G320799 | S1, S2, S7, S9, S3 | S1, S2, S7, S9 S3 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 |
| GRMZM2G110626 | S2, S4, S5, S1, S9 | S2, S4, S5, S1, S9 | S4, S5, S1, S9, S25, S16, S13, S15, S17, S14, S19, S26, S20, S12, S2, S24, S19 |
| GRMZM2G035131 | S7, S2, S5, S4, S1, S9 | S5, S4, S1, S9 | S4, S5, S9, S3, S1, S25, S19, S24, S17, S23 |
| GRMZM2G042080 | S4, S5, S2, S9, S3, S1 | S4, S5, S2, S9, S3, S1 | S4, S5, S9, S3, S1, S25, S19, S18, S2, S14, S15, S12, S16, S17, S26, S13 |
| GRMZM2G320799 | S1, S2, S7, S9, S3 | S1, S2, S7, S9 S3 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 |
| GRMZM2G315264 | S4, S5, S9, S1, S3, S2 | S4, S5, S9, S1, S3 | S9, S1, S3, S16, S17, S25, S18, S15, S26, S13, S5, S4 |
| GRMZM2G033626 | S1, S2, S16, S9, S5, S4 | S1, S2, S16, S9, S5, S4 | S1, S7, S2, S25, S16, S9, S3, S12, S17, S26, S13, S5, S4 |
| GRMZM2G320799 | S1, S2, S7, S9, S3 | S1, S2, S7, S9 S3 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 |
| GRMZM2G046558 | S2, S7, S1, S9, S3, S5, S4 | S2, S1, S5, S4 | S1, S2, S25, S17, S12, S26, S22, S13, S16, S6, S4, S5 |
|  | Most common: S1, S2, S9, S5, S4 | S1, S5, S4,( S9) | S25, S17, S1, S5, S4 (S16, S9, S13) |
| **Low score lncrnas L2** | **lncrna states** | **common** | **targ states** |
| L2 (GRMZM2G476477) | S4, S5, S7 | S4, S5 | S25, S17, S1, S5, S4 |
| GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 |
| L2 (GRMZM2G476477) | S4, S5, S7 | S4, S5 | S25, S17, S1, S5, S4 |
| L2 (GRMZM2G476477) | S4, S5, S7 | S4, S5 | S25, S17, S1, S5, S4 |
| GRMZM2G170336 | S2, S4, S5, S1, S9, S3, S16, S25, S13, S19 | S2, S4, S5, S1, S16, S25, S19 | S10, S2, S1, S5, S4, S25, S15, S16, S14, S17, S18, S23, S19, S12, S7 |
| GRMZM2G035131 | S7, S2, S5, S4, S1, S9 | S4, S5, S1, S9 | S4, S5, S9, S3, S1, S25, S19, S24, S17, S23 |
|  | Most common: S4, S5, (S7, S9, S1) | S4, S5, S1 | S25, S17, S1, S5, S4 (S7, S3, S2, S23, S19, S9) |
| **High score TG2** | **targ states** | **common** | **lncrna states** |
| TG2 (GRMZM2G041694 | S25, S17, S1, S5, S4 | S4, S5 | S4, S5, S7 |
| GRMZM2G094595 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 | S1, S2, S7, S9, S3 | S1, S2, S7, S9, S3 |
| GRMZM2G126128 | S4, S5, S9, S25, S17, S26, S13, S16, S1, S18, S12, S19, S23, S15, S24, S7, S2, S20, S22 | S5, S1, S9, S25, S2 | S5, S1, S9, S3, S25, S2 |
| GRMZM2G083732 | S4, S5, S1, S9, S25, S16, S13, S15, S17, S14, S19, S26, S20, S12, S2, S24, S19 | S2, S4, S5, S1, S9 | S2, S4, S5, S1, S9 |
| TG2 (GRMZM2G041694 | S25, S17, S1, S5, S4 | S4, S5 | S4, S5, S7 |
| GRMZM2G033519 | S25, S17, S13, S16, S12, S19, S24, S14, S1, S2, S4, S5 | S1, S2, S4, S5 | S9, S1, S2, S4, S5, S7 |
| GRMZM2G083732 | S4, S5, S1, S9, S25, S16, S13, S15, S17, S14, S19, S26, S20, S12, S2, S24, S19 | S2, S4, S5, S1, S9 | S2, S4, S5, S1, S9 |
| GRMZM2G126128 | S4, S5, S9, S25, S17, S26, S13, S16, S1, S18, S12, S19, S23, S15, S24, S7, S2, S20, S22 | S5, S1, S9, S25, S2 | S5, S1, S9, S3, S25, S2 |
| TG2 (GRMZM2G041694 | S25, S17, S1, S5, S4 | S4, S5 | S4, S5, S7 |
| GRMZM2G088291 | S17, S26, S25, S1, S9, S5, S4, S2 | S4, S5, S9, S2, S1 | S4, S5, S9, S2, S1 |
| GRMZM2G094595 | S25, S2, S9, S16, S18, S15, S1, S17, S24, S19, S23, S13, S20, S3, S5, S4, S7 | S1, S2, S7, S9, S3 | S1, S2, S7, S9, S3 |
|  | S25, S17, S1, S5, S4 (S2, S9, S16, S19, S24) | S5, S4, S9, S2 | Most common: S5, S9, S2 (S4) |
| **Low score TG2** | **targ states** | **common** | **lncrna states** |
| GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 |
| GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 |
| GRMZM2G035785 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 | S17, S25, S7, S4, S5, S9, S3, S1, S2 |
| GRMZM2G102483 | S10, S2, S1, S5, S4, S25, S15, S16, S14, S17, S18, S23, S19, S12, S7 | S2, S4, S5, S1, S16, S25, S19 | S2, S4, S5, S1, S9, S3, S16, S25, S13, S19 |
|  | S25, S17, S1, S5, S4, (S2, S7) | S1, S4, S5, S25, (S2) | S1, S2, S4, S5, S3, S25 |

Figure 14

This table represents a detailed chromatin state analysis by comparing those in the lncRNA subgroups and those in target gene subgroups. The blue highlights are the RNAs run for testing against FKBP4, the purple highlights are the RNAs run for testing against TAF15, and the green highlights are the RNAs run for testing against WDR43. There are several categories of data being analyzed, that considers the highest and lowest prediction scores from RPISeq, and compares the states present between the lncRNAs and the target genes. The chart considers states shared between L and TG subgroups. The yellow highlights indicate the most common states in either the L subgroup or the TG subgroup. The categories include: the highest and lowest scores from L1 and the chromatin state difference between the TG1 subgroup, the highest and lowest scores from TG1 and the chromatin state difference between the L1 subgroup, the highest and lowest scores from L2 and the chromatin state difference between the TG2 subgroup, and the highest and lowest scores from TG2 and the chromatin state difference between the L2 subgroup.

## Conclusion

The goal for this study was to determine whether changes in chromatin structure occurs, if these changes are due to the formation RBP-complexes, and if this could be considered a specific function of lncRNAs in maize. Though the output of lncRNA subgroup and RBP binding from RPISeq yielded inconclusive data, with no indication of any pattern specific to lncRNAs, it’s evident that chromatin structure is influenced by these interactions. However, the changing of chromatin is not a specific or primary function of lncRNAs in maize, rather the biproduct of numerous regulatory functions. Only considering the chromatin states of the genes targeted by lncRNAs and RBPs was not sufficient to identify specific target genes or the regulators of those genes. By considering the target gene chromatin states as well as the lncRNA chromatin states, detectable patterns arose. I was able to see the chromatin states and the specific epigenetic marks that operate in association with the states. There was a clear difference in the chromatin states and functions in the lncRNAs and the target genes. This difference appeared to act opposite of one another. If lncRNA had high binding affinity for a protein, the states present would allow for more expression or less condense chromatin through acetylation, while their corresponding target gene would have states in association with regulation or more condense chromatin, such as methylation or Mnase markers. For lower binding affinities in lncRNAs, the states would usually be those with less acetylation and their corresponding targets would have states that act to induce even more acetylation, preventing regulation. Overall, these interactions are extremely regulatory, but I don't think you can classify them by the chromatin states of their targets. It's simply to complex, with a lot of moving parts. I think we should study these interactions more to one day get a complete understanding, but I think the main finding from this project is that there are a lot of other variables to consider because the arrangement of chromatin state is something that arises as a result of several regulatory functions involved.

## Citations

*1.4. Support Vector Machines*. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/svm.html

Anthony, D. S.-B., & Hondermarck, H. (2023). *Transcriptomics*. Transcriptomics - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/transcriptomics

Batista, P. J., & Chang, H. Y. (2013, March 14). *Long noncoding RNAS: Cellular address codes in development and disease*. Cell. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3651923/

Bhattacharyya, N., Pandey, V., Bhattacharyya, M., & Dey, A. (2021, September). Regulatory role of long non coding RNAS (lncrnas) in neurological disorders: From novel biomarkers to promising therapeutic strategies. Asian journal of pharmaceutical sciences. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8609388/#:~:text=LncRNAs%20can%20regulate%20transcriptional%20processes,have%20post%20transcriptional%20regulatory%20functions.

Chatgpt. (n.d.). https://openai.com/chatgpt/

*Corn Genetics*. Southern Biological . (n.d.). https://www.southernbiological.com/corn-genetics/#:~:text=Corn%20is%20an%20excellent%20model,from%20the%20'parent'%20plants.

CRICK FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138-63. PMID: 13580867.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. arXiv.org. https://arxiv.org/abs/1810.04805v2

Eckardt, N. A., Birchler, J. A., & Meyers, B. C. (2022, July 4). *Focus on Plant Genetics: Celebrating Gregor Mendel's 200th Birth Anniversary*. The Plant cell. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9252500/#:~:text=Richardson%20and%20Hake%20point%20out,reviewed%20by%20Rhoades%2C%201984).

Ernst, J., & Kellis, M. (2012, February 28). *Chromhmm: Automating chromatin-state discovery and characterization*. Nature News. https://www.nature.com/articles/nmeth.1906

Fu, J., Cheng, Y., Linghu, J., Yang, X., Kang, L., Zhang, Z., Zhang, J., He, C., Du, X., Peng, Z., Wang, B., Zhai, L., Dai, C., Xu, J., Wang, W., Li, X., Zheng, J., Chen, L., Luo, L., … Wang, G. (2013, December 17). *RNA sequencing reveals the complex regulatory network in the maize kernel*. Nature News. https://www.nature.com/articles/ncomms3832

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., & Zavolan, M. (2021, March 4). *CLIP and complementary methods*. Nature News. https://www.nature.com/articles/s43586-021-00018-1

J. Huang a, a, AbstractEpigenetic gene regulation is important for proper development and gene expression in eukaryotes. Maize has a large and complex genome that includes abundant repetitive sequences which are frequently silenced by epigenetic mechanisms, Choi, Y., Haag, J. R., Jahnke, S., McGrath, J., Parkinson, S. E., Ream, T. S., Schläppi, M., Alleman, M., Barbour, J.-E. R., Barkan, A., Belele, C. L., Brink, R. A., Chandler, V. L., Chomet, P. S., Chopra, S., Coe, E. H., … Lisch, D. (2016, September 28). *Epigenetic control of gene expression in maize*. International Review of Cell and Molecular Biology. https://www.sciencedirect.com/science/article/abs/pii/S1937644816300715

Kaessmann, H. (2010, July 22). *Origins, evolution, and phenotypic impact of new genes*. CMU School of Computer Science. http://www.cs.cmu.edu/~durand/Phylogenetics/Readings/Kaessmann10.pdf

Kaikkonen, M. U., Lam, M. T. Y., & Glass, C. K. (2011, June 1). Non-coding RNAS as regulators of gene expression and epigenetics. Cardiovascular research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3096308/

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E., & Zdobnov, E. M. (2022b, November 9). *OrthoDB v11: Annotation of orthologs in the widest sampling of organismal diversity*. OUP Academic. https://academic.oup.com/nar/article/51/D1/D445/6814468

Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E., & Brendel, V. (2004, January 1). *MaizeGDB, the community database for maize genetics and genomics*. Nucleic acids research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308746/#:~:text=The%20Maize%20Genetics%20and%20Genomics%20Database%20(MaizeGDB%3B%20http%3A%2F%2Fwww,genetics%20and%20genomics%20of%20maize.

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., Yi, X., Xu, W., & Su, Z. (2018, January 4). *PCSD: A plant chromatin state database*. Nucleic acids research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753246/

MaizeGDB Gene Search Page. (n.d.). https://www.maizegdb.org/gene_center/gene#translate

Mercer, T. R., Dinger, M. E., & Mattick, J. S. (2009, March). *Long non-coding RNAS: Insights into functions*. Nature News. https://www.nature.com/articles/nrg2521

Muppirala, U. K., Honavar, V. G., & Dobbs, D. (2011, December 22). *Predicting RNA-protein interactions using only sequence information - BMC Bioinformatics*. BioMed Central. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-489

National Geographic. (n.d.). *Speciation*. National Geographic Education. https://education.nationalgeographic.org/resource/speciation/#

National Human Genome Research Institute. *A brief guide to genomics*. Genome.gov. (n.d.). https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics .

National Institutes of Health. (n.d.). *Homology: Orthologs and paralogs*. U.S. National Library of Medicine. https://www.nlm.nih.gov/ncbi/workshops/2023-08_BLAST_evol/ortho_para.html#:~:text=Homologous%20genes%20become%20separated%20in,duplication%20events%20are%20called%20paralogs.

NCI Dictionary of Genetics terms. National Cancer Institute. (n.d.). https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/gene#

Ong, Q., Nguyen, P., Thao, N. P., & Le, L. (2016, August). *Bioinformatics approach in plant genomic research*. Current genomics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4955030/

*ORTHODB presentation*. Prof Zdobnov laboratory. (n.d.). https://www.ezlab.org/orthodb.html

Phillips, T. & Shaw, K. (2008) Chromatin Remodeling in Eukaryotes. *Nature Education* 1(1):209. https://www.nature.com/scitable/topicpage/chromatin-remodeling-in-eukaryotes-1082/#:~:text=Interestingly%2C%20chromatin%20not%20only%20serves,proteins%20known%20as%20transcription%20factors.

Pranati, S., & Sarat, D. (2021). Reprogramming the Genome: CRISPR-Cas-based Human Disease Therapy. Non-Coding RNA - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/non-coding-rna

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., & Vandepoele, K. (2015, January). *Plaza 3.0: An access point for plant comparative genomics*. Nucleic acids research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384038/

RBPsuite. (n.d.). http://www.csbio.sjtu.edu.cn/bioinf/RBPsuite/

Shaikh, R. (2023, August 29). Mastering Bert: A comprehensive guide from beginner to advanced in natural language processing... Medium. https://medium.com/@shaikhrayyan123/a-comprehensive-guide-to-understanding-bert-from-beginners-to-advanced-2379699e2b51#:~:text=Introduction%3A,context%20and%20nuances%20in%20language.

State result. (n.d.). http://systemsbiology.cau.edu.cn/chromstates/search_state_result.php?state=3&species=Zm

Tabassum, H., & Parvez, S. (2021). *Translational Epigenetics in Neurodegenerative Diseases*. Chromatin Remodeling - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/chromatin-remodeling#:~:text=Chromatin%20remodeling%20is%20the%20rearrangement,DNA%20and%20control%20gene%20expression.

Tan, Y. C., Kumar, A. U., Wong, Y. P., & Ling, A. P. K. (2022b, July 15). *Bioinformatics approaches and applications in plant biotechnology*. Journal, genetic engineering & biotechnology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9287518/

U.S. National Library of Medicine. (n.d.). *FKBP4 FKBP prolyl isomerase 4 [Homo Sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/2288

U.S. National Library of Medicine. (n.d.-b). *Taf15 TATA-box binding protein associated factor 15 [homo sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/8148

U.S. National Library of Medicine. (n.d.-c). *WDR43 WD repeat domain 43 [mus musculus (house mouse)] - gene - NCBI*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/72515

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., … Yeo, G. W. (2020, July 29). A large-scale binding and functional map of human RNA-binding proteins. Nature News. https://www.nature.com/articles/s41586-020-2077-3

Wang, D. (2023, July 29). Biochemistry, RNA structure. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK558999/#:~:text=Ribonucleic%20acid%20(RNA)%20is%20a,guanine%2C%20uracil%2C%20and%20cytosine

Wang, Z., Gerstein, M., & Snyder, M. (2009, January). *RNA-seq: A revolutionary tool for transcriptomics*. Nature reviews. Genetics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/#:~:text=Understanding%20the%20transcriptome%20is%20essential,for%20understanding%20development%20and%20disease.

*What is Random Forest?*. IBM. (2021, October 20). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

Yamada, K., & Hamada, M. (2022, April 7). Prediction of RNA–protein interactions using a nucleotide language model. Academic.oup.com. https://academic.oup.com/bioinformaticsadvances/article/2/1/vbac023/6564689

Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., & Shen, F. (2019, November 8). *Mechanisms and functions of long non-coding RNAS at multiple regulatory levels*. International journal of molecular sciences. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6888083/#:~:text=lncRNAs%20are%20a%20new%20class,modification%2C%20primarily%20methylation%20and%20acetylation.

**Github Repository**

https://github.com/jtmckinley4/BSC-RESEARCH.git