Julian McKinley

Dr. Karen McGinnis

Spring 2024

Prospectus

The fundamental element of inheritance transmitted from one generation to the next is known as the gene. Genetic material in prokaryotes is typically minimal and housed in a single circular DNA molecule (Nature Education). In eukaryotes, genes consist of DNA sequences that are organized linearly along chromosomes within the cell nucleus. Genes serve as repositories of instructions directing the synthesis of proteins essential for expressing distinct physical attributes or traits, such as hair and eye color, as well as for performing specific cellular functions. (National Cancer Institute).

The entire set of genes or DNA in an organism is known 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, is concerned with 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 focusing on RNA molecules and their role in the transcriptome. By studying the transcriptome, we can better understand gene regulation and the genomic structure of organisms as a whole (Wang et al., 2009).

In human genetics, there has a been a massive genomic 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). Although there has been significant genomic research for humans, and thus mammals in general, there is much less research in genomics for plant genetics. Plant genomes are often larger and more complex than mammalian genomes, having high ploidy levels, repetitive sequences, and frequent genome rearrangements, which pose challenges for genome sequencing in plants. (Michael et al., 2015)

Despite these challenges, recent advances in genomic technologies have resulted in a rapid increase in sequence data (Tan et al). Next-Generation Sequencing (NGS) Technologies such as Illumina and PacBio, can uncover both DNA and RNA sequences. These technologies are cost effective and can investigate thousands of genes at once, enabling the sequences of DNA or RNA to be revealed (Thermo Fisher Scientific). Researchers are still at the forefront of assembling plant genomes, but with these advancements in sequencing, the potential for plant genomics has no limit. This potential is attributed to the fact that NGS sequencing not only directly improves genomic study in plants, but also that the genomic success for non-plant species indirectly correlates to plant genomic advancements. In other words, increasing genomic data for mammals can influence the success of plant genomics. We can see this with the accumulation of numerous databases such as PlantGDB, MaizeGDB, and MaizeDIG, which have sequence data available for various plant species including *Arabidopsis thaliana* and *Zea mays* (Tan et al., 2022).

Following this increase in accessible sequence data, researchers have sought to understand plant genomes and make genome-wide and transcriptome-wide identifications through comparative studies. While these studies can provide valuable insights into gene

conservation, evolution, and functional diversity across different plant species and between plants and other organisms, it's important to know why we should study plant genomes (Proost et al., 2015). Plant genomic research is important because they play vital roles for humans and various other organisms. They are not only regarded as sources of food and used for human medications, but they can be used as model organisms for studying regulation through epigenetic control (Ong et al., 2016). *Z. mays* serves as a great model for exploring the epigenetic control of gene expression due to its extensive genome rich in transposable elements, its demonstration of recognized forms of epigenetic gene expression regulation, and the numerous documented epigenetic phenomena available for in-depth studies (J. Huang a et al., 2016). Comparative genomics is an effective means for studying evolution, gene regulation, and gene functions. These studies compare species that have well-known genomes or transcriptomes such as humans, with species that lack genomic data such as *Z. mays* (National Human Genome Research Institute). By applying comparative studies, we can fill the knowledge gap in plant genomics and transcriptomics, and enhance our understanding of plant genomes in relation to human and mammalian genomes.

In summary, our understanding of plant genomes is relatively limited compared to human and mammalian genomes due to the larger size and complexity of plant genomes, as well as a lack of funding for research. As a result, there has been less genomic data in plant genetics in the past, however, NGS has allowed us to change this and discover more about the genomes and transcriptomes in plants. Using data from *Z. mays*, a comparative study that applies a transcriptomic approach, will take place to research the epigenetic regulation associated with the maize transcriptome.

The central dogma of molecular biology describes the flow of genetic information within a biological system. Proposed by Francis Crick in 1958, it states that genetic information 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).

RNA, or ribonucleic acid, is a polymer made up of nucleotides, whose primary function 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 (tRNA) 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 (rRNA) are responsible for the structure and catalytic activity of ribosomes, where protein synthesis ultimately 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).

In addition to the role for RNA described by the central dogma, research has also unveiled the regulatory roles of a type of RNA known as noncoding RNAs (ncRNAs). Making

up 98-99% of the human transcriptome, ncRNAs are RNA molecules that do not encode proteins but perform numerous functions in the cell that are crucial for gene regulation (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 known as long non-coding RNAs (lncRNAs) will be considered for this research project because of its epigenetic role in modulating gene expression, specifically at the chromatin level.

Typically exceeding 200 nucleotides in length, lncRNAs are found in both the nucleus and the cytoplasm (Kaikkonen et al., 2011). LncRNAs are noted for their epigenetic roles in modifying chromatin structure as well regulating gene expression both during transcription and after transcription has occurred. LncRNAs primarily form interactions with DNA, mRNA, and proteins (Zhang et al., 2019). Through these interactions, lncRNAs can control transcriptional processes by serving as chromatin remodelers or by altering histone proteins. They can also exert post-transcriptional regulatory roles and do so by regulating splicing, facilitating mRNA degradation, or inhibiting translation. In translation, they can serve as protein activity modulators, scaffolds, or as decoy receptors (Bhattacharyya et al., 2021).

Even though ncRNAs don't translate and undergo protein synthesis, they 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, researchers have only identified some ncRNA functions, specifically epigenetic regulation in lncRNAs, thus enduring difficulties understanding lncRNAs as a whole. These difficulties in understanding lncRNAs and how they

exert their regulatory functions, are due to their complex nature. Although research has uncovered a lot about their epigenetic roles in modulating gene expression, lncRNAs present several challenges in their study and understanding for several reasons. 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 ncRNAs. 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 rather a different means of conservation than protein-coding sequences. The massive diversity of functions in lncRNAs is another reason why they are difficult to predict. LncRNA may have unique roles in various cellular processes, making it hard to pinpoint specific functions associated with specific 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. Many lncRNAs are also 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 also be functional redundancy, where multiple lncRNAs may have overlapping or redundant roles, making it difficult to determine the specific function of individual lncRNAs (Mercer et al., 2009).

All these reasons not only highlight the difficulty in doing research on these molecules, but due to their lack of understanding, they have motivated myself and many other researchers to pursue this task to better understand the nature of lncRNAs and ncRNAs as a whole. In order to better understand lncRNAs, researchers have recognized the role of RNA-binding proteins (RBPs) and how they can provide valuable insights into the functions of lncRNAs.

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, influencing gene expression and cellular function (Hibah et al., 2022). Understanding the interactions of specific RBPs and specific lncRNAs can help give us insight on the types of regulation being applied. In other words, knowing what kind of RBP is involved, allows us to predict the regulatory functions of certain lncRNAs that interact with that particular RBP. For the purpose of this project, RBPs that assist in lncRNA epigenetic gene regulation through chromatin remodeling are of interest.

Chromatin remodeling refers to the rearrangement of chromatin into a variety of chromatin states by dynamically altering the placement of nucleosomes on DNA, implementing posttranslational modifications to histones, and/or through the methylation of cytosines in DNA, ultimately controlling gene expression. These chromatin states are arranged from compact states to states accessible by transcription, enabling replication, repair, and recombination (Tabassum & Parvez, 2021). *Z. mays* have 26 different chromatin states, each corresponding to different levels of regulation. Databases such as Plant Chromatin State Database (PCSD), can not only help us to understand the functions and epigenetic marks of chromatin states, but we can also map specific lncRNAs to specific chromatin states using tools like ChromHMM (Plant Chromatin State Data Base). This tool used to categorize genome data with their associated chromatin state, which have a particular chromatin mark. ChromHMM takes an input of gene annotations, then identifies and maps their particular chromatin state, revealing the kind of epigenetic regulation associated with each gene ID (Ernst & Kellis, 2012)

For the chromatin states in *Z. mays,* some of the epigenetic marks of these chromatin states include several types of histone methylations and histone acetylation's allowing or preventing

access to DNA. Other epigenetic marks associated with chromatin remodeling include nucleases, which catalyzes the breakdown of phosphodiester bonds, which hold DNA and RNA nucleotides together. Binding proteins can also be noted as epigenetic markers for chromatin remodeling (Plant Chromatin State Data Base). In particular, RBPs can bind to specific lncRNAs, that influence the remodeling of chromatin into specific 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).

Chromatin remodeling regulation via lncRNA and RBP interactions can be better understood using deep learning tools such as BERT-RBP. BERT-RBP can be used to model lncRNA and RBP interactions involved in chromatin remodeling and reveal the nature of these interactions in *Z. Mays.* BERT-RBP is based on BERT (Bidirectional Encoder Representations from Transformers) architecture, which is a deep learning language model created by Google (Shaikh, 2023). BERT-RBP incorporates the use of a pretrained model known as DNABERT, which is trained on the human genome and can be fine-tuned for specific tasks. BERT-RBP is fine-tuned to predict the binding affinities of RBPs and RNA molecules, and it requires the training of CLIP-seq (crosslinking immunoprecipitation) data. CLIP-seq data consists of the paired RNA sequences and RBPs allowing the BERT architecture to learn the complex relationships between the two (Yamada & Hamada, 2022). The BERT-RBP architecture takes an input RNA sequences which are split into 3-mers and tokenized, then converted into a 768-dimensional feature vector. These vectors undergo processing through 12 transformer encoder layers determining if the input RNA sequence binds to the RBP (Yamada & Hamada, 2022).

Experimental Plan

Since the emergence plant data availability, we have been able to uncover many unknowns of plant genomics and transcriptomics. We know that lncRNAs, which are present in plant transcriptomes, contribute to a variety of epigenetic regulatory functions, including chromatin remodeling. To better understand the nature of these functions, its best to analyze a specific function. So, lncRNAs that enable interactions with RBPs, causing gene regulation at the chromatin level, will be the basis for this research project to identify. BERT-RBP will be used to predict interactions between lncRNAs and RBPs associated with particular chromatin states. Since lncRNAs interact with certain RBPs that regulate gene expression through the remodeling of specific chromatin states, I'm going to predict which chromatin states will be associated with the interactions of lncRNAs in *Z. mays* and RBPs in humans. This is because we don't have CLIP-seq data for *Z. mays,* so I'm going to attempt to use a human model to make these predictions.

To begin, a .gtf file that consists of a list of lncRNAs and associated *Z. mays* gene IDs, obtained from the Nelson lab at Cornell, will be mapped to their associated chromatin states. This requires the data list to first be processed, which entails the removal of any duplicates as well as a gene ID conversion. The removal of duplicates is important because it gets rid of unnecessary data that will ultimately slow the program down. Then a gene ID conversion from ZM to GRMZM is necessary for the mapping process. This is because the Plant Chromatin State Database (PCSD), which will be used to map the chromatin states, takes inputs of GRMZM numbers, and the .gtf file obtained from the nelson lab consists of gene IDs in the ZM format. To make these conversions, Maize Genetics and Genomics Database (MaizeGDB) will be used.

Once duplicates are removed and converted into GRMZM numbers, batch searches of the data yielded, will be implemented in PCSD, revealing their associated chromatin states.

Once all lncRNAs are mapped and its understood how they are distributed across the chromatin states, only a few chromatin states will be narrowed down and considered for this project. Thus, the lncRNAs associated with those few chromatin states will only be considered. Then a list of human RBPs will be analyzed and narrowed down to a few RBPs associated with the particular chromatin states that were chosen.

Once the subgroups of lncRNAs and RBPs are determined through thorough analysis, BERT-RBP will be used to test the probability of these specific lncRNA and RBP interactions. BERT-RBP will model where they interact and how they interact. Based on the results an analysis will be made and questions like what preferences do the RBPs have, what about these lncRNAs causes this interaction, and how the results confirm or reject the interactions of lncRNAs and RBPs based on chromatin states? Overall, we will see if the *Z. mays* lncRNAs show the same predicted patterns of interactions with human RBPs as human lncRNAs. We can also follow up with additional analysis by looking at co-expression data by using SEEKR to compare motifs in human and plant lncRNAs.

Citations

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

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.

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

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

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

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/

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

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

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

Nature Education. (n.d.). Nature news. https://www.nature.com/scitable/definition/prokaryote-procariote-18/#:~:text=Most%20prokaryotes%20carry%20a%20small,surrounded%20by%20a%20nuclear%20membrane.

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/

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/

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/

ThermoFisher Scientific. (n.d.). *What is next-generation sequencing?: Thermo Fisher Scientific - US*. What is Next-Generation Sequencing? | Thermo Fisher Scientific - US. https://www.thermofisher.com/us/en/home/life-science/sequencing/sequencing-learning-center/next-generation-sequencing-information/ngs-basics/what-is-next-generation-sequencing.html#:~:text=Next%2Dgeneration%20sequencing%20(NGS),diseases%20or%20other%20biological%20phenomena.

Todd P Michael 1, 1, 2, Highlights•NGS speed and capacity enable over 100 published plant genomes.•Underserved specialty and orphan crop genomic resources grow due to low cost NGS.•Double haploid and diploid ancestors key to sequence complex plant genomes.•Polyploidy, The availability of plant reference genomes has ushered in a new era of crop genomics. More than 100 plant genomes have been sequenced since 2000, Albert, V. A., Initiative, A. G., Michael, T. P., Vogel, J. P., Rensing, S. A., Bennetzen, J. L., Zhang, G., Goff, S. A., Yu, J., Tuskan, G. A., Schnable, P. S., Schmutz, J., Consortium, T. G., Zimin, A., … Wu, J. (2015, February 19). *Progress, challenges and the future of Crop Genomes*. Current Opinion in Plant Biology. https://www.sciencedirect.com/science/article/abs/pii/S1369526615000229

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.

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%20

a%20new%20class,modification%2C%20primarily%20methylation%20and%20acetylation.