5-2022

Discovering the pathways and GO terms associated with Mettl3 modified circular RNAs in the embryonic cerebral cortex of mice.

Dunia Zedan
University of Louisville

DISCOVERING THE PATHWAYS AND GO TERMS ASSOCIATED WITH METTL3 MODIFIED CIRCULAR RNAs IN THE EMBRYONIC CEREBRAL CORTEX OF MICE

By

Dunia Zedan
B.S. in Biochemistry, University of Louisville, 2017

A Thesis Submitted to the Faculty of the
J.B. Speed School of Engineering
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science in Computer Science

Department of Computer Engineering and Computer Science
University of Louisville
Louisville, Kentucky

May 2022

DISCOVERING THE PATHWAYS AND GO TERMS ASSOCIATED WITH METTL3 MODIFIED CIRCULAR RNAs IN THE EMBRYONIC CEREBRAL CORTEX OF MICE

By

Dunia Zedan
B.S. in Biochemistry, University of Louisville, 2017

A Thesis Approved On

04/25/2022

By the following Thesis Committee:

_____
Juw W Park, Ph.D., Thesis Director

_____
Nihat Altiparmak , Ph.D

_____
Olfa Nasraoui, Ph.D.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Dr. Park who made this work possible. His valuable supervision and guidance supported me through all the stages of my thesis work.

I would like to thank the committee members Dr. Nasraoui and Dr. Altiparmak for their time and consideration.

Finally, I would be forever grateful for the support I had from my parents, my siblings, and my husband during my Masters' thesis journey.

ABSTRACT

DISCOVERING THE PATHWAYS AND GO TERMS ASSOCIATED WITH METTL3 MODIFIED CIRCULAR RNAs IN THE EMBRYONIC CEREBRAL CORTEX OF MICE

Dunia Zedan

April 25, 2022

Circular RNAs (cirRNAs) are a class of RNA molecules that result from the alternative back-splicing events that join the 3’ and 5’ ends normally present in the linear RNA molecules. It has been published that cirRNAs can function as gene regulators and as “microRNA sponges” to negatively control the functions of microRNAs. While many studies have been conducted to understand the regulatory roles of Mettl3 in linear messenger RNAs, fewer contributions were applied to understand the impact of Mettl3 modified cirRNAs on gene expression and on the regulation of different KEGG biological pathways and GO terms.

This thesis was conducted to identify the role of Mettl3 modification of cirRNAs in regulating gene expression and controlling different KEGG biological pathways and GO terms in the embryonic cerebral cortex of mice using high-throughput data sequencing. We constructed a generalized framework that led us to the identification of the cirRNA sequences that are significantly enriched in miRNA binding motifs and ultimately to the associated KEGG pathways and GO terms related to these interactions.

It has been found by this study that Mettl3 modification in cirRNAs can regulate gene expression by controlling different KEGG biological pathways and GO terms in a manner that is similar, but not identical, to their corresponding linear mRNAs. While some KEGG pathways and GO terms appeared to be regulated by the Mettl3 modification of both linear mRNAs and cirRNAs, few GO terms were regulated in mRNAs but not in cirRNAs. Interestingly, it has been found that Mettl3 modification in cirRNAs can promote the regulation of unique KEGG biological pathways and GO processes (not being regulated by the Mettl3 modified mRNAs) that are significant to the regulation of the neurological diseases' progressions such as brain tumors and intellectual disabilities in the embryonic cerebral cortex of mice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Circular RNAs (cirRNAs) are a class of non-coding RNA molecules that result from the alternative back-splicing events that are known as circularization. These back-splicing events generate single-strand covalently closed rings of RNAs that lack the 3’ and 5’ ends that normally present in linear RNA molecules. Lacking the polyadenylated tail (a long chain of adenine nucleotides) at the 3’ prime end prevents the degradation of cirRNAs by the RNA nuclei making them more stable and abundant compared to the linear RNAs. It has been found in many studies that cirRNAs can function as gene regulators and as microRNA (miRNAs) sponges to inhibit the functions of miRNAs. miRNAs are non-coding, single-stranded RNA segments with an average length of 22 nucleotides. miRNAs can bind to one or multiple sites in the linear mRNAs (messenger RNAs) and cirRNAs to regulate gene expression. The regulatory roles of miRNAs can be affected by the N6-methyladenosine (m6A) modification of mRNAs by increasing the availability of miRNA binding sites. While many studies have been conducted to understand the regulatory roles of m6A modification in linear mRNAs, fewer contributions were applied to understand the impact of miRNA interactions with m6A modified cirRNAs on gene expression and on the regulation of different biological pathways and processes.

1.1 Motivation

The different mammalians behaviors such as emotions, cognition, and the ability to move, to think, and to act are all controlled by the mammalian cerebral cortex. It has been established in multiple studies that N^6-methyladenosine (m^6A) modification, which is the

most important type of internal mRNA methylation, plays an important role in the regulation and development of the brain's functions by regulating the expression of genes and the translation of proteins. Three different types of enzymes control the modification of m^6A. The first type is the methyltransferase complex which controls the methylation of specific sites in the RNA transcripts. This type includes Mettl3, Mettl14, and Wtap enzymes. The second type is the m6A-binding proteins that are responsible for the promotion of messenger RNA (mRNA) translation by targeting specific m^6A sites. The last type of enzyme is demethylases such as Fto which are responsible for the extraction of the modification of m6A form the targeted RNA sequences [4].

A new research paper that was published on Jan 04, 2021, by a group of scientists aimed to understand the different biological functions of both Fto and Mettl3 in the embryonic development of the cerebral cortex of the mice's brains [4]. The study was conducted by the knockout (the deletion) of both Mettl3 and Fto in the mice embryonic cerebral cortex. The study concluded that the deletion of Mettl3 caused a fold in the structure of the cortex revealing the key important role of Mettl3 in the "proliferation and differentiation of neural progenitor cells" as suggested by the authors. Using both RNA_Seq (transcriptome sequence) and Ribo_Seq (ribosome profiling) high-throughput sequencing data of Fto and Mettl3 mRNAs knockout showed that both Fto and Mettl3 play a crucial role in the regulation of gene expression at both the transcriptional and translational stages of development. Mettl3 knockout, however, revealed that Mettl3 has a more significant role in "enhancing cell differentiation, neurogenesis, neuron differentiation, cell proliferation, cell cycle", as well as the efficiency of translation as concluded by the authors [4].

Starting from the important findings of the research paper above, the role of Mettl3 in regulating the gene expression and different biological pathways in the mice's embryonic cerebral cortex in the cirRNAs instead of the linear mRNAs needs to be discovered. Since cirRNAs are more stable, abundant, and conservative molecules compared to mRNAs, it is important to discover their entire role in regulating gene expression as well as in regulating different biological pathways and processes.

1.2 Thesis Contribution

To identify and understand the contributions of cirRNAs in regulating different biological pathways and processes, this thesis will focus on uncovering the additional/different Mettl3 regulation effects on the embryonic cerebral cortex development by studying the N6-methyladenosine (m6A) modification of cirRNAs instead of the linear RNAs. This study will focus on answering the following two important questions:

1. Does Mettl3 regulate the same biological pathways and processes in both linear mRNAs and cirRNAs?
2. Are there any unique biological pathways and processes initiated by the Mettl3 modification in cirRNAs that are not being detected in linear mRNAs?

Answering these two questions depends on following the developed and generalized cirRNA-miRNA framework. This framework starts with identifying and extracting cirRNA sequences from High-throughput data sequencing of the mice's embryonic cerebral cortex. This step will be followed by identifying the significant cirRNA sequences that are highly enriched in miRNAs binding motifs. The miRNAs that bind excessively to cirRNAs will be used to identify the different biological pathways and processes that get regulated by the miRNA-cirRNA interactions. Finally, the biological

pathways and processes that get regulated by the Mettl3 modified cirRNAs as discovered by our study will be compared with the proven regulatory functions of Mettl3 in the linear mRNAs of the mice's embryonic cerebral cortex.

1.3 Thesis outline

This thesis started with Chapter 1 that gives an introduction to the research topic and the questions that we need to answer. The rest of this thesis will be organized as follow. Chapter 2 will give a background and literature review of different biological and bioinformatics topics. Chapter 3 is our method section that will discuss the framework that we constructed and followed to find the significant miRNA-cirRNA interactions using high-throughput data sequencing. Chapter 4 presents our results and findings. Chapter 5 discusses our findings and their importance. The final chapter, Chapter 6, provides our conclusion and ideas for future works and developments.

CHPATER 2

BACKGROUND AND LITERATURE REVIEW

In this chapter, different biological and bioinformatics topics will be discussed. These topics will enhance the understanding of the entire research terms by giving all the required information and definitions of the different topics and terms used in this thesis. This chapter will start by giving background information about sequencing in general and how it was developed to the high-throughput sequencing methods that are widely used to sequence DNA and RNA molecules. A general discussion about transcription and translation processes will be followed. After that, cirRNAs and miRNAs will be defined and the different tools for detecting miRNA-cirRNA interactions will be outlined. Finally, the KEGG pathway and GO terms will be discussed.

2.1 Overview of sequencing

The genome is the genetic material that includes all the genetic information of an organism. Genome sequencing is the fundamental key for all studies involving the understanding of the genetic material. By sequencing the genome, the exact composition and order of the nucleotides within the genome can be identified [6]. Sequencing a specific gene can provide important information about the function of the protein being encoded by the gene or about the related gene regulation events. Since the early decades of the discovery of the DNA structure by Watson and Crick in 1953, great efforts have been applied to develop sequencing methods and techniques that can facilitate the process of DNA and RNA sequencing. The first successful attempt of sequencing was in 1965 when

the sequencing of a 76 bases long alanine tRNA-RNA was established by Robert Holley and colleagues [9]. Following the tRNA sequencing attempt, the first successful sequencing of a protein-coding gene (bacteriophage MS2 coat protein) was established in 1972 by Walter Fier and colleagues [9].

Since the mid-1970s, the purification and sequencing of segments of a few bases of DNA were achieved. In 1975, Frederick Sanger and Walter Gilbert developed the 'plus and minus' sequencing technique that was based on the synthesis of nucleotides by using a single-stranded DNA template, a primer, and a DNA polymerase. Accordingly, the first DNA genome sequencing was in 1977 in which the 5368bp of phage φX174 was sequenced by Sanger and colleagues. In the same year 1977, Sanger improved his method and developed the chain-termination method also known as the Sanger method. This method depends on using dideoxynucleoside (ddNTPs) that lacks the 3′-hydroxyl group to terminate the elongation process of segments and generate segments of different lengths. This method opened the door to the development and enhancement of the first-generation sequencing methods that include shotgun sequencing, recombinant DNA (rDNA) technologies, and polymerase chain reaction (PCR) [9].

2.2 First Generation Sequencing

While the sequencing of a single gene can provide considerable information about the function of the protein encoded by a gene or the regulation function of a specific gene, the sequencing of the entire genome can reveal considerable information about how the genes that form an organism are directed to provide the required growth and development during different stages of the organism's life. Furthermore, comparing the sequenced genes of different individuals can help with studying mutational events that can be associated with a particular disease, it can identify inherited disorders among individuals,

and help with understanding the characterization, development, and progression of many diseases [7]. The human genome project was initiated in 1990 to generate the entire DNA sequence of the human genome. This project cost around three billion dollars and took around 15 years to be fully completed in 2003 [6].

The contribution of the human genome project in understanding the nature of the human genome has motivated the researchers and scientists to develop faster, more affordable, and more effective approaches to genome sequencing and led to the development of the second-generation sequencing known as High-throughput next-generation sequencing (HT-NGS) technologies [7].

2.3 High-throughput data sequencing (HT-NGS)

The great accomplishment of the human genome project derived the scientist to apply tremendous efforts in developing more advanced sequencing techniques. These techniques are known as the High-throughput next-generation data sequencing (HT-NGS). All the developed methods follow the same approach of preparing a library from a DNA molecule that is either native or being amplified. The process starts with generating DNA fragments and identifying a particular fragment size. These fragments are covalently bound to a solid surface. Next, the ligation of adapters at the ends of each generated fragment is performed. DNA amplification by the polymerase chain reaction (PCR) is then applied and followed by a series of sequencing reactions that utilize the use of the amplified copies to capture fluorescently labeled deoxynucleotides to elongate the DNA segments [6]. The generated sequences are then assigned a Phred quality score (Q) which is being used to identify the accuracy of the generated sequence. Phred quality score follows the following equation: Q = $-10log_{10}P$ and it is used to identify the probability (P) of the incorrect assignment of a base within the sequence [7].

HT-NGS focuses on advancing and generating new sequencing methods that are affordable and capable of generating complete genome sequencing in a limited time frame. To store the tremendous amount of genomic data, databases were developed and enhanced by algorithmic functions to deposit and retrieve the stored information to be used for data mining and bioinformatics analysis and research. In addition, HT-NGS played a role in other applications including whole-exome sequencing, chromatin immunoprecipitation sequencing, the determination of the genome-wide epigenetic landscape, and the high-throughput RNA sequencing (RNA-Seq) and ribosomal sequencing (Ribo-seq) [6].

2.4 RNA-Sequencing (RNA-Seq)

High-throughput second-generation data sequencing has been applied to discover the transcriptome profiling of cells. Transcriptomes refer to the transcription of all genes within a cell including the full set of messenger RNAs (mRNAs) and the non-coding RNAs under a specific set of conditions and during different stages of cell development. The recently developed high-throughput method for transcriptome profiling is RNA sequencing (RNA-Seq). This method generates high-quality results related to the quantification of gene expression, the generation of novel transcripts, the detection of the variations in intron splicing, as well as the discovery of the expressions that are allele-specific [12].

High-throughput RNA sequencing process starts with the isolation of a population of RNA molecules to remove the undesired ribosomal RNAs (rRNAs) by either using the mRNAs' poly(A) enrichment method or the rRNA depletion methods. While the poly(A) enrichment method that depends on using oligo-dT primers that bind to the mRNAs with poly-A tails is a very effective method for rRNA depletion, it is not capable of recognizing the mRNAs with missing poly(A) tails that result due to the degradation of the sample or

the formation of circular RNAs. The rRNA depletion method depends on using specialized probes or primers that can selectively recognize rRNA and deplete them [22]. The mRNAs are then converted to a library of copied DNA (cDNA). Following the previous step, an adapter is attached to one or both ends of the generated cDNAs. A sequencing step is then followed to generate short reads from either one end of the fragments generating single-end reads or from both ends of generating paired ends reads [7]. If a reference genome is available, the reads are aligned to the reference genome. Alternatively, if a reference genome is not available, the overlapping reads are being aligned against each other and be assembled de novo to generate a genomic transcriptional map [11].

2.5 DNA transcription and translation

Transcription is the process of converting a portion of the DNA sequence into an RNA sequence known as the messenger RNA (mRNA). In transcription, the double helix of the DNA molecules is being opened forming two single-stranded DNA sequences and the nucleotides of each DNA strand become exposed. A molecule called RNA polymerase will bind to one of the DNA strands to initiate the transcription [10]. The selected DNA strand will act as a templet to generate the reverse-complementary RNA sequence by the base pairings of nucleotides. The matching bases from DNA to RNA are A base-pairs with U, T base-pairs with A, C base-pairs with G and G base-pairs with C.

Translation, on the other hand, is the process of converting the mRNA sequence into a linear chain of amino acids or proteins. To generate the proteins, a biological structure called the ribosome is responsible for translating the mRNA sequence such that each codon (a set of three nucleotides) forms one amino acid according to the codon table shown in Figure 2.3 [3].

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	STOP	UGA	STOP	A
	UUG	Leu	UCG	Ser	UAG	STOP	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	CAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	CAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	CAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	CAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Figure 2.1: The codon table for the formation of amino acids [3].

2.6 Detection of circular RNAs (CirRNAs)

In 1990, a new class of RNA molecules was identified and referred to as circular RNAs (cirRNAs). cirRNAs are non-coding RNA molecules that result from the alternative back-splicing events known as circularization. There are two known types of back-splicing that can generate multiple cirRNAs from a single gene locus. The first type is called the alternative 5’ back-splicing and it is a reversed orientation that involves the alternative bindings of two or more 5’ back-splicing downstream sites to the same 3’ back-splicing upstream site. The other type which is called the 3’ back-splicing is the opposite of the first type such that it involves a reversed orientation that alternatively links two or more 3’ back-splicing upstream sites to the same 5’ back-splicing downstream site [24].

Published studies and research had shown the existence of cirRNA isoforms for a thousand of genes within human cells. It was also found that the cirRNA isoforms are more abundant and more conserved than the compatible liner messenger RNAs (mRNA) from the same gene due to lacking the poly-A tail that is usually present at the 3’ end of mRNAs [20]. The earlier discovered role for cirRNAs is their ability to function as microRNA sponges or modulators. In addition, many studies have suggested the relationship between specific diseases and the existence of a high number of cirRNAs in the cancer tissues, blood, saliva, and other cellular tissues [23]. Some of the studies focused

on cirRNA that are directly related to a specific medical condition by being present in one condition only and absent in the other which are characterized as tissue-specific or disease-specific. Interestingly, cirRNAs can also be found in both conditions of normal tissues and abnormal tissues but different quantities [2]. The above-discovered roles of cirRNAs led to an increase in the efforts applied to discover their existence and roles within different tissues and at different conditions.

The advances in HT-NGS technologies enabled scientists to develop scientific tools that help with the detection of cirRNAs by using RNA-seq data. Some of the developed tools and software to detect cirRNAs are CIRCexplorer, CIRI, MapSplice, find_circ, DCC, UROBORUS, KNIFE, circRNA_finder, and NCLscan. These tools were developed to detect the back-spliced junction sites that generate the cirRNA reads and are being mapped to two unrelated proteins within the genome. For this study, the newly developed seekCRIT method for cirRNA detection will be applied. SeekCRIT (seek for differentially expressed CircRNAs In Transcriptome) is a computational tool developed in October 2020 for the detection of the differently expressed cirRNAs within tissues using rRNA depleted RNA-seq data [2]. The three needed components that seekCRIT uses to detect cirRNAs are the RNA-seq data in a FASTQ format, the entire genome sequence in a Fasta format, and the transcriptome in GTF or RefSeq formats. seekCRIT is characterized by its ability to work with replicates and the use of a junction-count-based estimation approach to generate a normalized quantification of cirRNAs [2].

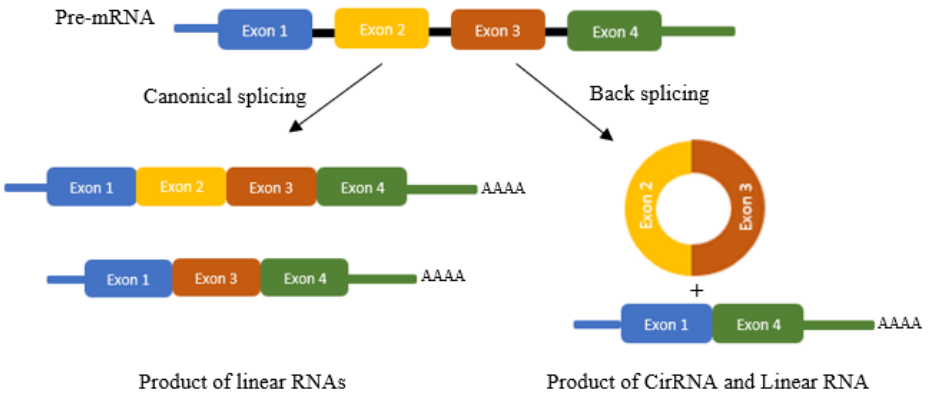

Figure 2.2: The formation of cirRNAs by back-splicing [2]

2.7 Overview of miRNAs

The first discovery of miRNAs goes back to 30 years ago when the miRNA lin-4 was identified in the nematode Caenorhabditis elegans by Ambros and Ruvkun in 1993 [15]. It was found that lin-4 RNA coded for regulatory RNA of 22 nucleotides in length that did not code for any protein [8]. Since then, major studies have been conducted to discover miRNAs and their roles. It has been found that miRNA exists in most eukaryotes. It was discovered that miRNAs account for 1% to 5% of the entire human genome and that they have regulatory functions over more than 30% of the protein-coding genes of the human genome [14].

miRNAs are non-coding, single-stranded RNA segments with an average length of 22 nucleotides. A series of biological events are involved in transcribing DNA sequences to produce precursor miRNAs by RNA polymerase II and RNA polymerase III that then get converted to active and mature miRNAs through multiple cleavage actions [8]. miRNAs play an important role in gene regulation at both the transcription and translation levels. This function is achieved by loading the miRNAs into the effector complex RNA

induced silencing complex (RISC) that targets a specific messenger RNA (mRNA) through base pairing to regulate its expression directly and negatively and/or repress translation [18].

The regulation of mRNA is achieved by the binding of the seed sequence (motif) of miRNAs to the 3’ UTRs of the target mRNAs. The seed sequence is the first 2–8 bases on the 5′ end of miRNAs and they bind to mRNAs by base-pairing [14]. For cirRNAs, miRNAs bind to the coding DNA sequences region (CDS) since cirRNAs lack the 3’ UTR region. The function of cirRNAs as miRNA sponges is related to their interactions with one or multiple miRNA binding sites and thus gene expression can be regulated by inhibiting the activity of miRNAs.

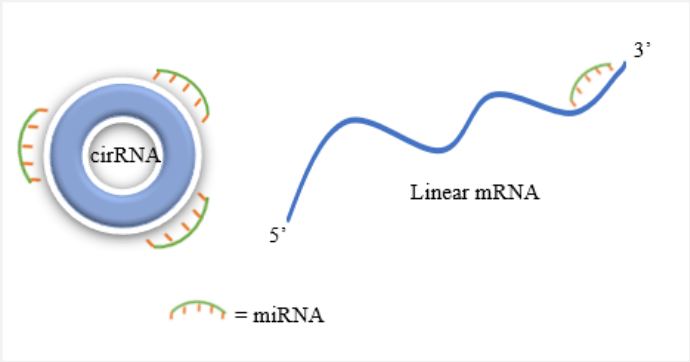

Figure 2.3: “miRNA sponge” activity of cirRNAs by the existence of multiple binding sites [16]

2.8 MicroRNAs targeting prediction tools

Since the first discovery of miRNAs, researchers started to focus on developing computational tools and algorithms that can predict the special interactions between miRNAs and mRNAs sequences. These developed tools follow the most important approach which is seed pairing that has been applied for many years. The seed pairing approach considers the Watson–Crick base-pair interactions of the miRNAs’ seed

sequences (the first seven nucleotides from the 5' prime region) with the reverse-complementary mRNA sequences [14]. Some of the widely used bioinformatics tools for miRNAs target predictions include TargetScan, miRanda, miRDB, PicTar, miRIAD, and DIANA-mirPath. Although these tools have shown successful and accurate miRNAs target predictions, they usually require providing the species gene's symbol, or Ensembl gene ID, or transcript as an input to make the predictions by finding the binding sites on the 3' prime UTR region of the sequence where the miRNAs usually bind. Since our research focuses on generating the cirRNAs sequences first that are mainly formed from the CDS region before detecting the miRNAs targets, we generated our generalized code that works on finding the miRNA-cirRNA interactions for any species given the cirRNA and miRNA sequences in Fasta format using the seed-pairing approach.

2.9 KEGG pathway and GO category analysis

KEGG stands for Kyoto Encyclopedia of Genes and Genomes, and it is an online database that combines all the biological pathways including all molecular and genomic information, diseases, and drug information to facilitate the understanding and the interpretations of different bioinformatics research results [13]. The different biological pathways that form the KEGG include the cell and its intracellular compartments, the organism, and interactions with the ecosystem. This database was first constructed in May 1995 by the Kanehisa Laboratories under the Human Genome Program and many development actions have been conducted across the years to create the current KEGG dataset that can be used to integrate and interpret huge amounts of sequencing data generated by high-throughput sequencing techniques [16]. To better understand the functions of miRNAs, KEGG will be used in this research to spot the biological pathways that are regulated by the miRNAs in question.

GO refers to Gene Ontology which is a dataset that combines more than 38,000 biological vocabularies that describe three different aspects of biological events. These biological events are the molecular level functions that result from the gene's product, the cellular components, and anatomy, as well as the biological processes of the cellular activities [1]. Running GO analysis on miRNA sequences will help with identifying which molecular functions, cellular components, and/or biological processes are being regulated by the miRNAs in question.

For both KEGG pathways and GO categories, mirPath v.3 which is an online software specialized in the detection of the miRNA's regulation functions will be used. This tool was reported to generate accurate results for different biological research involving the detection of the functional activities of miRNAs.

CHPATER 3

FRAMEWORK FOR THE DETECTION OF MICRORNA - CIRCULAR RNA INTERACTIONS

Chapter 3 outlines the methods and framework that we developed to identify the significant miRNA-cirRNA interactions in the Mettl3 RNA-seq data. This chapter starts with the overall developed pipeline that will take us from the RNA-seq data (from this section) to the KEGG pathway and GO terms (in the results and discussion section). Following the pipeline, the data collection method will be discussed. The coordinates of cirRNAs are generated in the next step by using seekCRIT program. DNA sequences (that will be converted to RNA sequences) and the miRNA sequences of Mus musculus organism (mouse) are extracted from different databases. Finally, our developed miRNA-cirRNA code will be introduced and its functions in identifying the cirRNAs that are highly enriched in miRNA binding motifs will be discussed.

3.1 Research Overview and the proposed pipeline

The goal of this thesis is to identify the roles of the Mettl3 modified cirRNAs in the regulation of gene expression and the associated biological KEGG pathways and GO processes. These roles will be identified by finding the cirRNAs that are highly enriched in miRNA binding motifs which will uncover the regulatory effect of Mettl3 on the miRNA-cirRNA interactions.

The following two questions were introduced in Chapter 1:

1. Does Mettl3 regulates the same biological pathways and processes in both linear mRNAs and cirRNAs?
2. Are there any unique biological pathways and processes initiated by the Mettl3 modification in cirRNAs that not being detected in linear mRNAs?

To answer these questions, the following pipeline will be followed:

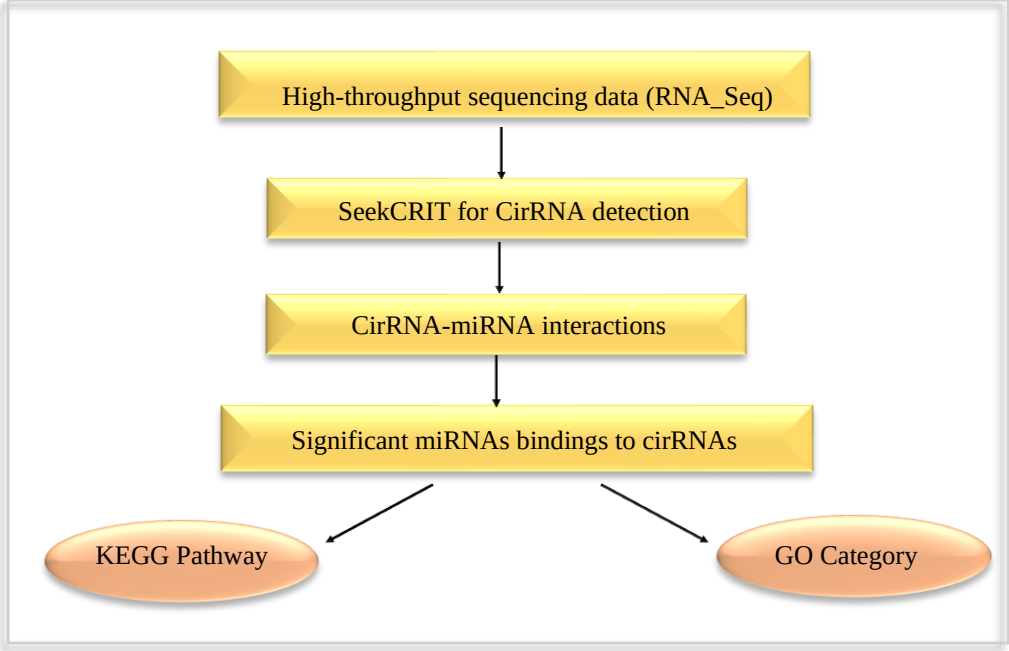

Figure 3.1: The research's pipeline framework

3.2 High-throughput sequencing data collection

The RNA_Seq data collected in the study of the Mettl3 regulations of embryonic cerebral cortex development in mice outlined above was used as an input for the seekCRIT program to generate cirRNA sequences. Two groups of RNA_Seq with one replicate for each group were used. One wild type of RNA_Seq replicate (with no modifications, used as a control) and one RNA_Seq replicate of Mettl3 knockout were downloaded in FASTQ file formats. The two RNA_seq files are:

- RNA_Seq wild type (GSM4692796 RNA_WT1) and knockout Mettl3 (GSM4692799 RNA_ME1) FASTQ sequences.

Using both a wild type of RNA_seq and the Mettl3 knockout RNA_seq with seekCRIT will generate cirRNA sequences that are common in both biological conditions. This in turn will be essential in finding the biological conditions that are regulated/controlled by Mettl3 due to lacking its sequence in the generated cirRNAs.

3.3 Generating cirRNAs using seekCRIT

In addition to the RNA_Seq data downloaded above, SeekCRIT tool used the following packages and input files to generate the circular RNAs running under the Linux system environment:

Packaged instilled for seekCRIT:

- STAR Aligner: v2.5.2b
- pysam >=0.9.1.4
- numpy >=1.11.2
- scipy
- fisher (Fisher's exact test)
- mne (FDR calculation)

Files downloaded for seekCRIT:

All genome files downloaded from the UCSC Genome Browser:

1. Mouse genome (GRCm38/mm10) gtf annotation file: mm10.refGene.gtf.gz
2. Mouse genome (GRCm38/mm10) FASTA file: mm10.fa.gz.
3. Mouse genome (GRCm38/mm10) refseq formatted file: mm10.ref.txt.

seekCRIT utilizes the files outlined above to generate cirRNAs sequences with the following associated information: the chromosome in which the cirRNA formed, cirRNA 5' end position, cirRNA 3' end position, DNA strand (+/-), number of exons within the cirRNA, size of exons within the cirRNA, offsets of exons included in the cirRNA, circRNA type (circRNA, ciRNA, ccRNA), gene's name, isoform's name, Index exons or intron, flanking introns (Left intron/Right intron), read count of the circular junction in

first and second samples, read count of the linear junction in first sample and second samples, percent back splicing index (PBI) for first and second samples, difference between PBI values of two samples, pValue, and FDR.

Following the steps/commands outlined in the seekCRIT website [21], the program was run the required files and sequences outlined above to generate cirRNA sequences. The program generated a total of 433 cirRNA distributed across the entire range of mouse chromosomes. These cirRNA are differ in many aspects including the strands in which they were formed, their lengths, and their exons' count and size.

a

chrom	circRNA_start	circRNA_end	strand	exonCount	exonSizes	exonOffsets	circType	geneName	isoformName	exonIndexOrIntronIndex
chr12	61669479	61669654	+	1	175,	0,	circRNA	Lrfn5	NM_178714	2
chr9	78480241	78480418	-	1	177,	0,	circRNA	Eef1a1	NM_010106	6
chr6	137426765	137441140	+	3	35,82,88,	0,3906,14287,	circRNA	Ptpro	NM_011216	18,19,20
chr6	6006269	6010580	+	2	141,126,	0,4185,	circRNA	Dync1i1	NM_001191027	14,15
chr6	124932108	124934003	+	4	68,127,36,50,	0,302,668,1845,	circRNA	M1f2	NM_145385	2,3,4,5
chr6	103708714	103708906	+	1	192,	0,	circRNA	Ch11	NM_007697	21
chr6	93792346	93815825	-	2	207,120,	0,23359,	circRNA	Magi1	NM_001286786	19,20
chr6	22965161	22987457	+	6	96,67,158,151,185,127,	0,565,7550,7850,20969,22169,	circRNA	Ptprz1	NM_001081306	5,6,7,8,9,10
chr6	115034782	115034952	-	1	170,	0,	ciRNA	Tamm41	NM_026894	6
chr6	51464094	51464547	-	2	117,120,	0,333,	circRNA	Hnrnpa2b1	NM_016806	3,4
chr6	148294514	148325306	-	3	168,122,190,	0,11483,30602,	circRNA	Tmtc1	NM_198967	11,12,13
chr6	148180823	148180983	+	1	160,	0,	ciRNA	Far2	NM_178797	12
chr6	72957912	72958382	-	2	112,46,	0,424,	circRNA	Tmsb10	NM_001190327	2,3
chr6	34598670	34599106	+	1	436,	0,	ciRNA	Cald1	NM_001347100	1
chr3	90087456	90087562	+	1	106,	0,	circRNA	Tpm3	NM_001293749	3

b

FlankingIntrons	CircularJunctionCount_Sample_1	LinearJunctionCount_Sample_1	CircularJunctionCount_Sample_2	LinearJunctionCount_Sample_2	PBI_Sample_1
chr12:61524830-61669479|61669654-61839407	143	1	67	1	0.044444444444444446
chr9:78480148-78480240|78480420-78480510	1	32936	1	44458	
chr6:137420407-137426764|137441143-137442644	1	110	0	0	
chr6:5972209-6006269|6010580-6027371	1393	0	0.005063291139240506	0.0	
chr6:124931586-124932098|124934007-124934290	1	2967	0	0	
chr6:103708537-103708707|103708912-103709122	2249	0	0	0.015810276679841896	
chr6:93785646-93792346|93815825-939432031	80	0	0	0.024390243902439025	
chr6:22961747-22965161|22987457-22994517	1	78	0	0	
chr6:115032312-115034950	1	121	0	0	
chr6:51463493-51464088|51464547-514651781	8348	0	0	0.00023952095808383233	
chr6:148284970-148294514|148325306-148335642	1	13	0	0	
chr6:148175148-148180985	1	0	0	0	
chr6:72957652-72957912|72958385-72958578	1	29112	0	0	
chr6:34598665-34662127	18	0	0	0.2	
chr3:90086579-90087454|90087572-90087673	11440	0	0	0.0013869625520110957	

c

PBI_Sample_2	deltaPBI(PBI_1-PBI_2)	pValue	FDR	
0.028985507246376812	0.015458937198067634	0.6466695919420935	1.0	GT-AG
6.072014087072682e-05	4.498425555105713e-05	1.5735885360155552e-05	0.9999999999982432	AT-CC
0.017857142857142856	0.0	0.017857142857142856	1.0	CG-GG
0.005063291139240506	1.0	1.0	1.0	GT-AG
0.00067362748400134 72	0.0	0.00067362748400134 72	1.0	CA-CC
0.0	0.015810276679841896	1.0	1.0	AA-TT
0.0	0.024390243902439025	1.0	1.0	CT-AC
0.025	0.0	0.025	1.0	GT-AG
0.016260162601626018	0.0	0.016260162601626018	1.0	TC-AT
0.0	0.00023952095808383233	1.0	1.0	CT-AC
0.13333333333333333	0.0	0.13333333333333333	1.0	CT-AC
1.0	0.0	1.0	1.0	AG-GC
6.869547296833139e-05	0.0	6.869547296833139e-05	1.0	TT-AC
0.0	0.2	1.0	1.0	AG-AG
0.0	0.0013869625520110957	1.0	1.0	GT-AG

Figure 3.2: Sample output from seekCRIT program (all three parts a, b, and c are one continues output)

3.4 Collecting DNA sequences

The seekCRIT output provide complete information about the exons that form each cirRNA. Using the first seven columns of the seekCRIT output (chrom, circRNA_start,

circRNA_end, strand, exonCount, exonSizes, and exonOffsets), the sequences of each exon were extracted and downloaded in a Fasta format. This was achieved by firstly converting the seekCRIT output file to an Excel file in a csv (comma separated values) format.

Table 3.1: Converting seekCRIT output to a csv Excel sheet

chrom	starting_base	exon_start	exon_end	strand	exonCount	exonCount -1	exonSize
chr12	61669479	61669479	61669654	+	1	0	175
chr9	78480241	78480241	78480418	-	1	0	177
chr6	137426765	137426765	137426800	+	3	2	35
chr6	137426765	137430671	137430753	+	3	2	82
chr6	137426765	137441052	137441140	+	3	2	88
chr6	6006269	6006269	6006410	+	2	1	141
chr6	6006269	6010454	6010580	+	2	1	126

As shown in Table 3.1, the cirRNA_start and cirRNA_end columns were changed to exon_start and exon_end, respectfully. This step was necessary to define the exact sequence of each exon within each cirRNA. In addition, two extra columns were added. The first column is called starting_base (added as a second column), and it contains the same starting value for the exons that form a single cirRNA. The second additional column is exonCount-1 (the seventh column) which was used to generate duplicate rows for each of the exons that form a single cirRNA sequence. The duplication step of rows was achieved by using the Excel Kutools feature that enables copying the data based on the desired columns value, in this case, the exonCount-1 column. Table 3.2 below is an illustrative example of using Kutools to duplicate rows.

Table 3.2: Duplication of rows using Kutools illustration example

Chrom	Count	Count-1	Chrom	Count	Count-1
chr1	1	0	chr1	1	0
			chr2	2	1
chr2	2	1	chr2	2	1
			chr3	3	2
chr3	3	2	chr3	3	2
			chr3	3	2

To apply Kutools, the data columns were highlighted in the Excel sheet, and the following options were selected: Kutools > Insert > Duplicate Rows/Columns based on cell value. The duplication was established based on the Count-1 column which generates the dataset with duplicate rows.

Next, based on the exonOffsets values and the exonSize columns, the exon_start and exon_end bases were calculated and modified to generate the sample results shown in Table 3.1 such that the start of each exon equals to its associated starting_base+exonOffset and the end of each exon equals to its starting_base+exonOffset+exonSize. For example, the data from the third row up to the fifth row generate one cirRNA that starts at base 137426765 and ends at base 137441052 and this cirRNA has three exons with their starting and ending bases.

After generating the datasets that show the starting and ending bases of the exons, the three columns: chrom, exon_start, and exon_end were used to extract the exact DNA sequences using SeqTailor webserver. SeqTailor is an efficient and rapid approach for the extraction of both DNA and protein sequences from many species including human, mouse, rat, cow, zebrafish, and others. This web server can generate the DNA sequences based on multiple options selected by the user. These options include the reference genome, the strand, the window size for displaying the output, and requires the input of the

chromosomes' values in BED format which is the format that utilizes each feature in one line with three required fields: the chromosome, the starting base, and the ending base. The output will be generated in a Fasta format.

As an input to the SeqTailor webserver, the reference genome was set to mouse [mus musculus] (GRCm38), the strand was set to forward (positive strand), the chrom, exon_start, and exon_end columns from the Excel sheet output were used.

Table 3.3: Sample input information into SeqTailor software in BED format

chrom	exon_start	exon_end
chr12	61669479	61669654
chr9	78480241	78480418
chr6	137426765	137426800
chr6	137430671	137430753
chr6	137441052	137441140
chr6	6006269	6006410

```
>12_61669479_61669654|+
GAGCTCCTGTATCCATTCAGCCGGGTATTGAGAAGATATGTAGTAGCTGACAACCTTCTGTCTGACAGCACTCTCAGAAC
CAGCATTGGAAATGCGTTCACCCTGCCGGTCTTCTGCAGCAGCCAGGCCACGTGCTGACTCCTTCACGGTAGTAAGGAAC
TCTCCAGGCTTCAGAA

>9_78480241_78480418|+
CTGGGATGTGCCTGTAATCATGTTTTTGATGAAGTCTCTGTGTCCTGGGGCATCAATGATGGTCACATAGTATTTGCTGG
TCTCGAATTTCCACAGGGAGATGTCAATAGTGATACCACGCTCACGCTCAGCTTTCAGTTTGTCTAAGACCCAGGCGTAC
TTGAAGGAGCCCTTTCCC

>6_137426765_137426800|+
GAGTAAAAATGGCTTAAAGAAGAGGAAACTAACAAA

>6_137430671_137430753|+
GCCCCGTTCAGCTGGATGACTTCGATTCTTACATCAAGGATATGGCCAAGGACTCGGACTATAAATTCTCTCTTCAGTTT
GAG

>6_137441052_137441140|+
GGAGTTGAAGTTGATTGGACTGGATATTCCGCACTTTGCTGCAGATCTACCGCTGAACCGATGTAAAAACCGCTACACAA
ACATCCTGC
```

Figure 3.3: SeqTailor output DNA sequences in Fasta format

3.5 Collecting the mouse miRNA sequences

After the previous step of getting all DNA sequences of the exons that form the generated cirRNAs, the mouse miRNA sequences were downloaded. The miRNA dataset

of all species was downloaded from the miRBase which is an open-source biological database that contains all published miRNA sequences and annotations. This database contains microRNA sequences for up to 271 organisms with human, mouse, rat, worm, and fly being the most popular. It was recorded in 2019 that miRBase has a total of 38589 hairpin precursors and 48860 mature microRNAs [13]. All mature microRNA sequences have been downloaded from miRBase in a Fasta format. The required set of microRNAs for the mouse species will be filtered later when used with the miRNA_cirRNA code.

```
>mmu-miR-15b-5p MIMAT0000124 Mus musculus miR-15b-5p
UAGCAGCACAUCAUGGUUUACA
>mmu-miR-15b-3p MIMAT0004521 Mus musculus miR-15b-3p
CGAAUCAUUAUUUGCUGCUCUA

>mmu-miR-15a-5p MIMAT0000526 Mus musculus miR-15a-5p
UAGCAGCACAUAAUGGUUUGUG
>mmu-miR-15a-3p MIMAT0004624 Mus musculus miR-15a-3p
CAGGCCAUACUGUGCUGCCUCA
```

Figure 3.4: microRNA naming convention and sequences form miRBase

For naming/identifier meanings, please consider the first row: mmu-miR-15b-5p MIMAT0000124. In this row, the first three letters (mmu) specify the organism (Mus musculus), the miR symbols show that these are mature miRNAs, the number 15 is a sequential numbering of miRNAs based on the time they get published, the letter suffix (b) that follows the number 15 shows that this mature miRNA sequence is closely related to mmu-miR-15a-3p that have the same name/identifier, but different suffix (a). The 5p and 3p that follows the letter suffix represent the direction of the miRNA (either from the 5'arm or the 3' arm). The MIMAT0000124 refers to the miRBase Accession number.

3.6 miRNA-cirRNA to identify the micro-RNA-circular-RNA binding sites

After collecting all needed DNA and micoRNA sequences, a program called "miRNA-cirRNA" has been developed to discover the binding sites between microRNAs and cirRNAs as well as to calculate the occurrence of the multiple interactions that could occur between each pair of miRNA and cirRNA sequences. The generated code focuses on spotting all seven-mer interactions between cirRNAs and miRNAs. The seven-mer seed sequence is the sequence that starts from base 2 through base 8 in the miRNA sequences. The miRNA-cirRNA code goes through the following steps:

- Uploading all mature miRNA sequences (Fasta format).
    - Extract only the Mus musculus miRNA sequences: A total of 1636 mature miRNA sequences were found.
    - Convert all sequences to be read from the 3' prime to 5' prime
- Upload all DNA exons sequences.
- Concatenate the exon sequences that belong to the same cirRNA sequence and include all the associated information from seekCRIT (chrom, circRNA_start, circRNA_end, strand, exonCount, exonSizes, and exonOffsets).
- Based on the strand sign, convert DNA sequences to RNA sequences through transcription:

```
If strand = '+':
        RNA = Transcription of DNA
else:   #strand = '-'
        RC = Reverse-Complement of DNA
        RNA = Transcription of RC
```

- Since cirRNAs are in the form of a circler loop, concatenate the last 6 bases to the beginning of the RNA sequences and concatenate the first 6 bases to the end of the

RNA sequences. This will ensure finding the binding sites that occur at the bases where the splicing event occurred to join the 5′ splice site with the 3′ splice site.

- Generate the 7-mer seed sequences from each miRNA (5’ base 2 to base 8).
- Find the complement sequences (complement-seed) for each of the generated 7-mer seed sequences
- Find all matching pairs of the complement-seed and the cirRNAs:

```
save = 0
end_base = 0
cirRNA_list = []
complement_list = []
for i in all ciRNA sequences:
        for j in all complement-seed:
                find all matching pairs(j, i)
                        if (i == save and end_base <= j.end()):
                                continue                #eliminate overlapping
sequences
                        else:
                            cirRNA_list.append(i)
                            complement_list.append(j)
        save = i
        end_base = j.end()
```

- Find the repeated interactions between the same complement-seed and cirRNA pair by grouping the results based on these two columns.
- Find the length of each cirRNA sequence.
- The final output of the program will generate the following seven columns: seekCRIT cirRNA information, miRNA sequence, Seed, Complement seed, cirRNA sequence, cirRNA length, Number of binding sites.

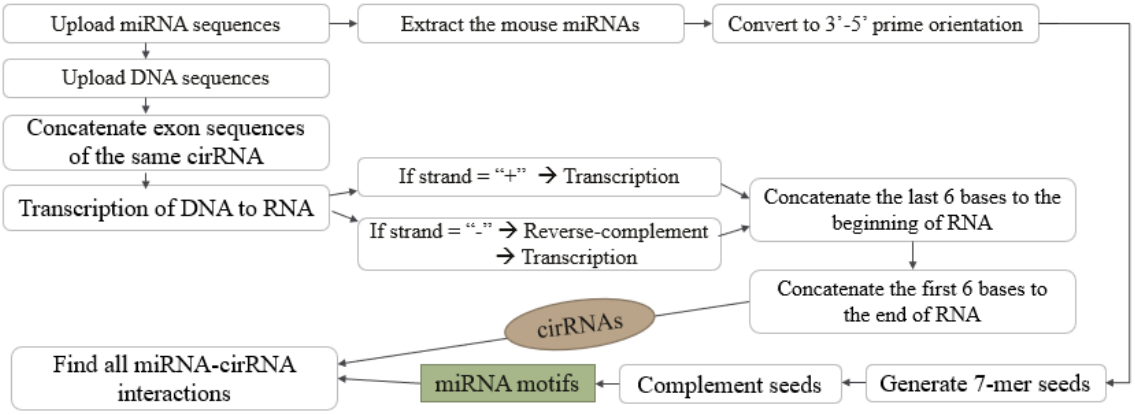

Figure 3.5: miRNA-cirRNA code workflow

seekCRIT cirRNA information (chrom	circRNA_start	circRNA_end	strand	exonCount	exonSizes	exonOffsets)	miRNA sequence	Seed	Complement seed	cirRNA sequence	cirRNA length	Number of binding sites
chr11	95572040	95574181	-	1	2141,	0	UAUACAUACACGCAC	ACACAUA	UGUGUAU	AUCCAUCAACGUAGC	2141	5
chr10	86704293	86705439	-	2	103,133,	0,1013	ACUGCCCUAAGUGCU	UCCUUCU	AGGAAGA	GGAAGGUUUCCCAG	236	4
chr2	140181573	140187782	+	3	48,104,39,	0,2324,6170	ACUGAGUAGAGGUA	AGGAGGA	UCCUCCU	UCCCAGCACUAUUAG	191	3
chr6	148180823	148180983	+	1	160,	0	UACGAGUGCGAGUG	GGGACGG	CCCUGCC	CCCUGCCUGUGUCU	173	2
chr11	95572040	95574181	-	1	2141,	0	CGGCGGUGGUGGUG	GUGGUGU	CACCACA	AUCCAUCAACGUAGC	2141	12
chr2	18203604	18203745	+	1	141,	0	GACGGACUCGUGGG	AGGGACA	UCCCUGU	GGGACUCCCUGUGA	154	2
chr2	160753447	160753737	+	2	117,95,	0,195	GCUCUAGACUUCGG	UCCCGUU	AGGGCAA	GGAGCCGAAGCUGG	226	1

Figure 3.6: miRNA-cirRNA code partial output

CHAPTER 4

RESULTS

Running the miRNA-cirRNA code on our 433 cirRNAs and 1636 miRNAs generates a total of 10386 7-mer seed reverse-complementary interactions between the cirRNAs and miRNAs. To find the cirRNA sequences that are significantly enriched in miRNAs binding motifs, statistical analysis methods were conducted. Following the statistical analysis, KEGG pathway and GO category analysis were conducted.

4.1 P-value using Fisher exact test

The first statistical measure is the probability value (p-value). A p-value is the measure of the probability of generating results that are at least as extreme/large as the obtained results that supports the null hypothesis. More specifically, the p-value will help identify the likeness of generating random results. P-value has a range from 0 to 1 and as the p-value decreases, the results become more significant and stronger evidence will be available to reject the null hypothesis. P-values will be calculated using Fisher's exact test. Fisher exact test is being used to determine if there is no random relationship between two independent categorical variables in a 2x2 contingency table [5]. For the sample table below (Table 4.1), Fisher's exact test can be calculated using the following formula:

$$\text{P-value} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!}$$

Table 4.1: The 2x2 contingency table to calculate p-values using Fisher exact test [5]

	Column 1	**Column 2**	**Total**
Row 1	a	b	a + b
Row 2	c	d	c + d
Total	a + c	b + d	a + b + c + d

Where ‘a’, ‘b’, ‘c’, and ‘d’ are the frequencies of the categorical variables that create the 2x2 contingency table.

To run Fisher's exact test on our data, the following 2x2 contingency table was computed:

Table 4.2: The 2x2 contingency table to find the p-values for miRNA-cirRNA interactions

	Exact Binding Sites	Possible – Exact Binding Sites
Real Data	A	B
Expected Data	C	D

Where:

- Exact Binding Sites column refers to the number of binding sites for each miRNA-cirRNA interaction.
- Possible – Exact Binding Sites column refers to all the possible (theoretical) 7-mer seed interactions that can occur between the miRNA and the cirRNA in each row minus the obtained exact value of binding sites.
- Real Data row refers to the actual data that was obtained by running the code.
- Expected Data row represents the theoretical data that can be extracted from the Mouse’s genome.

A, B, C, and D values can be extracted/calculated as follow:

A: 7-mer seed interactions extracted from the seventh column of the code's output.

B: This value is calculated using the following formula:

$$\textit{Possible number of binding sites} = \frac{Length\ of\ the\ motif}{Length\ of\ cirRNA} - \textit{Exact Binding Sites}$$

- Such that the length of the motif is always equal to 7 (the 7-mer seed) and the length of the cirRNA can be extracted from the sixth column of the code's output.

C: This value represents the frequency of finding each specific motif within the entire Mouse's genome (mm10.fa file).

- Linux terminal is used to find the frequency of each motif. For example, to find the occurrence of ACACACA motif within the mm10.fa genomic file, the command below was executed and the value of 287758 was obtained:

  pcregrep -M ACACACA mm10.fa | wc -l

- C can then be calculated using the following formula:

  $$C = \frac{The\ frequency\ of\ the\ Motif}{\left(\frac{the\ size\ of\ the\ genome}{Motif\ size}\right)}$$

  Where the size of the Mouse's genome is 2.5 billion DNA letters long.

D: Similar to B, this value is calculated by subtracting the Exact number of binding sites of the expected data from the possible number of binding sites.

After generating the 2x2 contingency table for each row of the miRNA-cirRNA code, Python code was used to run the Fisher Exact test on the entire data. To run Fisher Exact test on python, scipy library and the stats module need to be installed and called, respectfully.

Python to run the Fisher Exact test:

```
import scipy.stats as stats
p-values = []
for i in the list of contingency_tables:        #i is a 2x2 contingency table
        p-values.append(stats.fisher_exact(i)))
print(p-values)
```

Based on the p-values that were obtained by running the Fisher Exact test on all the 2x2 contingency tables, the dataset was arranged in ascending order (from smallest to largest) as shown below.

Tracking Number	seekCRIT cirRNA information (chrom	circRNA_start	circRNA_end	strand	exonCount	exonSizes	exonOffsets)	miRNA sequence	Seed	Compliment_seed	cirRNA sequence	cirRNA length
1	chr11	95572040	95574181	+	1	2141,	0	ACGCACGCACACACAC	ACACACA	UGUGUGU	UGUAUGUGGUGUGUC	2141
2	chr11	95572040	95574181	-	1	2141,	0	AUGUAAGGAAGUGUC	GUGUGUG	CACACAC	GCUGUAAAUGAUCAGA	2141
3	chr11	95572040	95574181	-	1	2141,	0	UACAUAAAGAAGUAU	GUAUGUG	CAUACAC	GCUGUAAAUGAUCAGA	2141
4	chr11	95572040	95574181	+	1	2141,	0	UACAUACACGCACACAU	ACACAUA	UGUGUAU	UGUAUGUGGUGUGUC	2141
5	chr11	95572040	95574181	+	1	2141,	0	UACACACACACAUACAC	AUACACA	UAUGUGU	UGUAUGUGGUGUGUC	2141
6	chr11	95572040	95574181	-	1	2141,	0	AUGUACAUGUGUGUC	GUGUGUA	CACACAU	GCUGUAAAUGAUCAGA	2141
7	chr11	95572040	95574181	-	1	2141,	0	ACGUGUACGUGUGUAC	UGUAUGU	ACAUACA	GCUGUAAAUGAUCAGA	2141
8	chr11	95572040	95574181	-	1	2141,	0	GCGGUGGUGGUGGUC	GUGGUGU	CACCACA	GCUGUAAAUGAUCAGA	2141
9	chr11	95572040	95574181	+	1	2141,	0	CACUUCUCUUCCCACCA	CACCACA	GUGGUGU	UGUAUGUGGUGUGUC	2141
10	chr11	95572040	95574181	-	1	2141,	0	UGAGUGUGUGUGUGU	GUGUGAG	CACACUC	GCUGUAAAUGAUCAGA	2141
11	chr11	95572040	95574181	+	1	2141,	0	GGCCUGUACUCUCACA	CUCACAC	GAGUGUG	UGUAUGUGGUGUGUC	2141
12	chr8	93967089	93967413	-	1	324,	0	ACAUGUACAUGCGUGU	CGUGUGU	GCACACA	UUAGAAAAAUGUUUUU	324
13	chr8	93967089	93967413	-	1	324,	0	AUGUAAGGAAGUGUC	GUGUGUG	CACACAC	UUAGAAAAAUGUUUUU	324
14	chr12	29811382	29811730	-	1	348,	0	CCCUCACCCUUCCUCCU	CUCCUUC	GAGGAAG	UCUGAAGACAAUGAUC	348
15	chr12	29811382	29811730	-	1	348,	0	GCCCUAAGUGCUCCUU	UCCUUCU	AGGAAGA	UCUGAAGACAAUGAUC	348
16	chr11	95572040	95574181	-	1	2141,	0	UACAUACACGCACACAU	ACACAUA	UGUGUAU	GCUGUAAAUGAUCAGA	2141
17	chr11	95572040	95574181	-	1	2141,	0	ACAUGUACAUGCGUGU	CGUGUGU	GCACACA	GCUGUAAAUGAUCAGA	2141
18	chr11	95572040	95574181	-	1	2141,	0	GUGCUUCCACUUUGUC	UUGUGUG	AACACAC	GCUGUAAAUGAUCAGA	2141
19	chr10	79865631	79865783	-	1	152,	0	GCCCUAAGUGCUCCUU	UCCUUCU	AGGAAGA	GGCCCCAGCAGAGCAGC	152
20	chr10	79865631	79865783	-	1	152,	0	CCCUCACCCUUCCUCCU	CUCCUUC	GAGGAAG	GGCCCCAGCAGAGCAGC	152

Figure 4.1: Top 20 cirRNAs highly enriched in miRNA binding motifs

Table 54.3: The associated information for the 20 top miRNA-cirRNA interactions

Tracking Number	N of binding sites	Motif size	Possible binding sites	N Possible – N binding	N expected binding	N possible binding – expected	p-value
1	43	7	306	263	0	306	0
2	39	7	306	267	0	306	0
3	21	7	306	285	0	306	0
4	19	7	306	287	0	306	0
5	16	7	306	290	0	306	0
6	15	7	306	291	0	306	0
7	14	7	306	292	0	306	0
8	12	7	306	294	0	306	0

9	12	7	306	294	0	306	0
10	7	7	306	299	0	306	0.015
11	7	7	306	299	0	306	0.015
12	5	7	46	41	0	46	0.057
13	5	7	46	41	0	46	0.057
14	5	7	50	45	0	50	0.058
15	5	7	50	45	0	50	0.058
16	5	7	306	301	0	306	0.062
17	5	7	306	301	0	306	0.062
18	5	7	306	301	0	306	0.062
19	4	7	22	18	0	22	0.109
20	4	7	22	18	0	22	0.109

It has been observed that the p-values were distributed across the entire range from 0 to 1. Importantly, a direct negative relationship between the number of binding sites and the p-values was observed such that increasing the number of binding sites led to a decrease in associated p-values. The length of the cirRNAs also plays a role in the resulted p-values. For the same number of binding sites, the shorter cirRNAs have lower p-values compared to the longer cirRNAs. These are the two most important factors that affect the p-values and hence the significant measure of each miRNA-cirRNA reverse-complement interaction. After organizing the dataset based on p-values, the top 20 miRNA-cirRNA interactions were selected. The table below shows the obtained results.

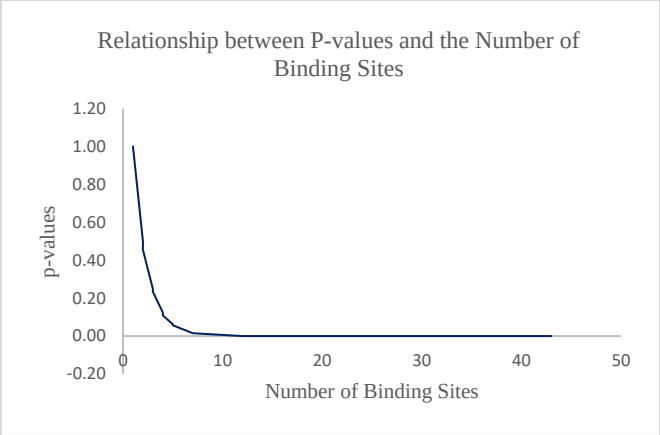

Figure 4.2: The relationship between p-values and the number of binding sites

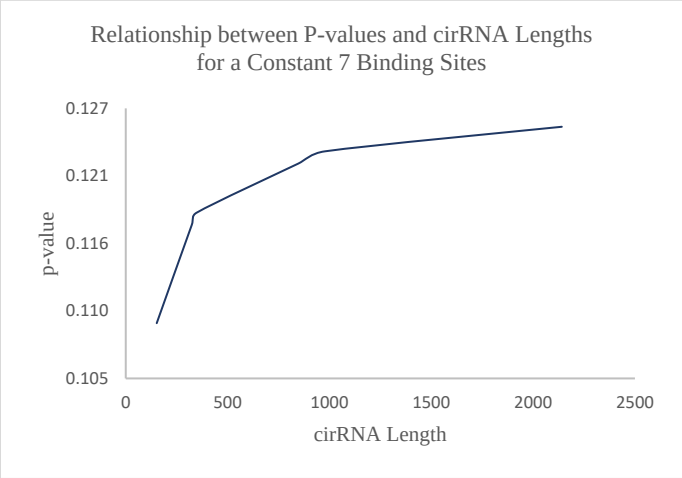

Figure 4.3: The relationship between the length of cirRNA and p-values for the same number of binding sites

From table 4.2, the first nine miRNA-cirRNA interactions have a p-value close to 0. This indicates that these interactions are highly significant, and these events did not occur randomly (by chance). Similarly, the miRNA-cirRNA interactions in rows 10 and 11 have a p-value of $0.015 < 0.05$ indicating the lower significance of these interactions. Starting from row 12, the p-values started to overcome the significant p-value cutoff of 0.05 resulting in no-significant interactions. As a result, the null hypothesis (no

relationship/interactions between miRNA-cirRNA) cannot be rejected. Accordingly, only the first 11 rows will be used for further statistical measures and analysis.

4.2 False Discovery Rate (FDR) Calculations

The second statistical measure that will be used to identify the significance of interactions is the False Discovery Rate (FDR). FDR is a statistical measure that is used to test the accuracy of multiple test comparisons. This approach is widely used to correct the false significant measures that appear significant due to random events. FDR is applied for the significant corrections of many experiments, especially, the high-throughput sequencing experiments. When calculating the p-values for a large number of experiments for a single test, the chance of false-positive results for the entire test is high and it can be measured using the following formula: FB = $1 - (1 - \sigma)^m$, where σ is the significant level of 0.05 and m is the total number of tests performed (miRNA-cirRNA interactions) [19]. Accordingly, FDR measure is needed to correct the p-values by eliminating the false-positive results.

FDR can be calculated by using the following the steps outlined below:

1. The ordered p-values (from lowest to largest) will be ranked. An additional Rank column has been added to the dataset.
2. The FDR at σ significant level of 0.05 is calculated using the following formula:

$$\text{FDR} = \left(\frac{\text{Rank}}{\text{Total number of tests}}\right) * \sigma$$

3. P-values will then be compared with the FDR values for each test. The tests with p-values<= FDR will pass and consider significant.

Accordingly, for the resulted 11 significant miRNA-cirRNA interactions, an additional ranking column was added. FDR values were calculated for all these rows using

σ significant level of 0.05 and total number of tests of 10386. The following table has been generated:

Table 4.4: The resulted FDR values for the 11 significant cirRNA-miRNA interactions

Tracking Number	p-value	miRNA names (Top 11)	Rank	FDR
1	0	mmu-miR-466h-3p MIMAT0017274 Mus musculus miR-466h-3p	1	0
2	0	mmu-miR-206-3p MIMAT0000239 Mus musculus miR-206-3p	2	0
3	0	mmu-miR-1b-3p MIMAT0017326 Mus musculus miR-1b-3p	3	0.0001
4	0	mmu-miR-466d-3p MIMAT0004931 Mus musculus miR-466d-3p	4	0.0001
5	0	mmu-miR-466f-3p MIMAT0004882 Mus musculus miR-466f-3p	5	0.0001
6	0	mmu-miR-466a-5p MIMAT0004759 Mus musculus miR-466a-5p	6	0.0001
7	0	mmu-miR-297a-5p MIMAT0000375 Mus musculus miR-297a-5p	7	0.0002
8	0	mmu-miR-7044-5p MIMAT0027992 Mus musculus miR-7044-5p	8	0.0002
9	0	mmu-miR-6917-3p MIMAT0027735 Mus musculus miR-6917-3p	9	0.0002
10	0.015	mmu-miR-574-5p MIMAT0004893 Mus musculus miR-574-5p	10	0.0002
11	0.015	mmu-miR-12189-3p MIMAT0049848 Mus musculus miR-12189-3p	11	0.0003

From Table 4.4 above and by comparing the p-values to their corresponding FDR values, only the first nine rows show p-values that are lower than or equal to the corresponding FDR values. As a result, these nine columns show significant reverse-complementary interactions between miRNAs and cirRNAs. These top 9 cirRNAs that are significantly enriched in miRNAs binding sites will be used to find the associated biological pathways and categories (KEGG pathway and GO category).

After finding the 9 cirRNAs that are significantly enriched in miRNAs binding motifs, the miRNAs that are associated with each cirRNA-miRNA interaction will be extracted. The table below shows the 9 miRNAs involved in these interactions (the table is organized in an ascending order based on p-values as well as FDR values):

Table 4.5: The final 9 miRNAs that have significant interactions with cirRNAs

TRACKING NUMBER	MIRNA NAMES (TOP 9)
1	mmu-miR-466h-3p MIMAT0017274 Mus musculus miR-466h-3p
2	mmu-miR-206-3p MIMAT0000239 Mus musculus miR-206-3p
3	mmu-miR-1b-3p MIMAT0017326 Mus musculus miR-1b-3p
4	mmu-miR-466d-3p MIMAT0004931 Mus musculus miR-466d-3p
5	mmu-miR-466f-3p MIMAT0004882 Mus musculus miR-466f-3p
6	mmu-miR-466a-5p MIMAT0004759 Mus musculus miR-466a-5p
7	mmu-miR-297a-5p MIMAT0000375 Mus musculus miR-297a-5p
8	mmu-miR-7044-5p MIMAT0027992 Mus musculus miR-7044-5p
9	mmu-miR-6917-3p MIMAT0027735 Mus musculus miR-6917-3p

4.3 KEGG pathway and GO category using mirPath

The mirPath v.3 online software was used to find the biological pathways and the biological processes that are regulated by the nine miRNAs extracted above. The mirPath v.3 takes species and miRNAs as input and generates KEGG pathways and GO terms as output. The nine miRNAs whose motif sequences were enriched in the detected cirRNAs were provided as an input. KEGG and GO analysis contains the following four columns. The first column gives the detected KEGG pathways or GO categories, the second column gives the enrichment p-values (using σ significant level of 0.05), the third column gives the number of genes that have been regulated, and the last column gives the number of miRNAs (from the input miRNAs) that can regulate genes in each specific path/process.

# GO Category	p-value	#genes	#miRNAs
1. cell (GO:0005623)	<1e-325	1361	7
2. intracellular (GO:0005622)	1.09034170674e-232	1259	7
3. biological_process (GO:0008150)	3.90022133994e-64	1743	7
4. cell differentiation (GO:0030154)	3.09213741441e-62	424	6
5. ion binding (GO:0043167)	1.6582470844e-45	694	7

Figure 4.4: The output table format of the GO Category using mirPath v.3

4.3.1 KEGG pathway analysis

It has been found that the 9 miRNAs are involved in the regulation of 25 different KEGG biological pathways. These pathways involve the regulation of fatty acid biosynthesis and metabolism, N-Glycan biosynthesis, and cancer initiation and progression including Glioma (tumor in the brain and spinal cord), thyroid cancer, leukemia, melanoma (skin cancer), and colorectal cancer. In addition, other pathways regulated by the extracted miRNAs include the endocytosis process (when cells engulf and absorb external material through their cell membrane), Dorso-ventral axis formation (how the embryonic axes/shape is performed during the earlier stages of development), signaling different pathways including the foxO (proteins that play an important role in regulating metabolism, and cellular proliferation, and repairing DNA), GnRH (a hormone made by the hypothalamus part of the brain the regulates the reproductive functions), thyroid hormones (hormones that regulate the function of the thyroid gland), and MAPK (the pathway that regulates cellular proliferation, differentiation, apoptosis as well as inflammatory responses). Moreover, these miRNAs can regulate the mRNA surveillance pathway (the mechanism that works on the degradation of abnormal mRNA sequences), adherence junction (the cell-cell adhesion complexes that help the cells to respond to biomedical signal and changes in structures and plays an important role in the development of embryos), and gap junction (the channels that connect the adjacent cells and play an important role in the regulation of phycological process).

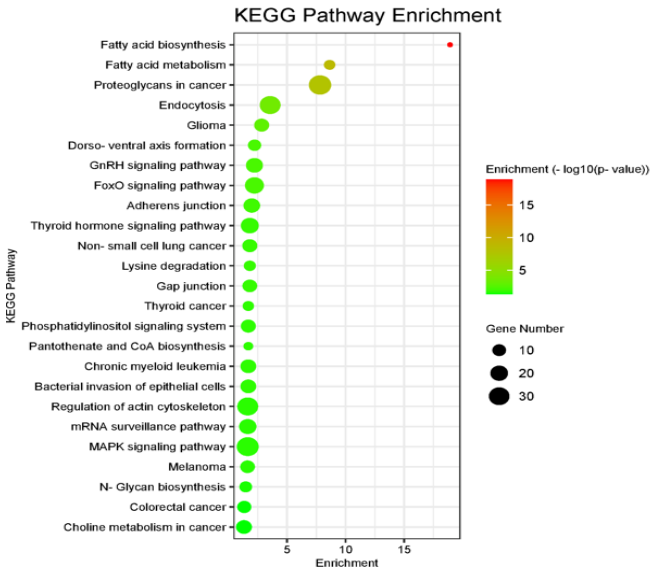

Figure 4.5: Bubble map for KEGG Pathway enrichment analysis of miRNAs. The y-axis identifies the KEGG pathway, the x-axis defines the enrichment (-log10(p-value)), -log10(p-value) are represent by the color scale, and the number of genes is represented by the size of the nodes.

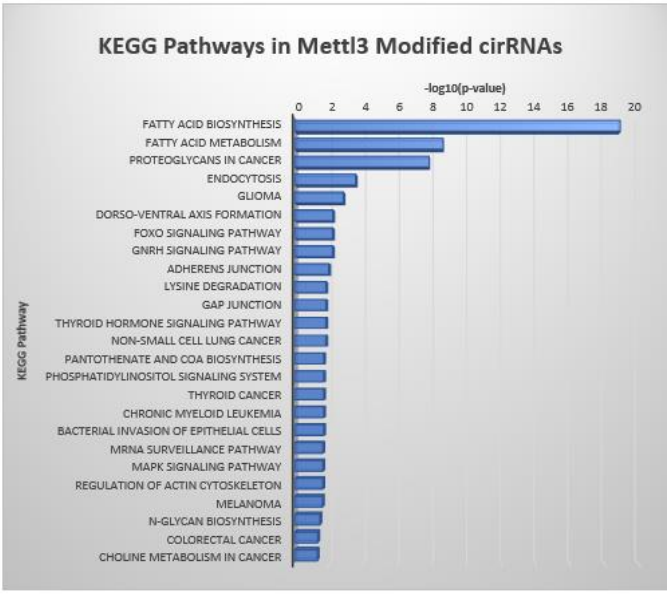

Figure 4.6: KEGG pathway analysis of miRNAs. The y-axis displays the KEGG pathway, and the x-axis displays the -log10(p-value).

4.3.2 GO category analysis

The same 9 miRNAs undergo GO analysis, and 34 biological processes appear to be regulated by these miRNAs. These biological processes included the cell, intracellular, anatomical structure development, ion binding, molecular function, embryo development, cell division, cell differentiation, cellular protein modifications, cellular nitrogen compound metabolic process, chromosome organization and nuclear chromosome, biosynthetic process, cytoplasm, cell morphogenesis, cytoskeleton organization, cell motility, catabolic processes, cytoskeleton, homeostatic process, Golgi apparatus, and developmental maturation. All of these GO terms are essential for the development and maturation of the embryonic cerebral cortex of mice.

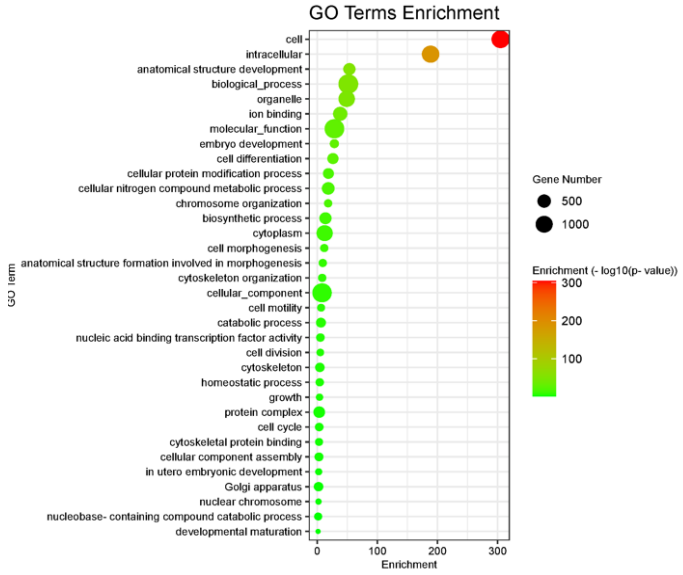

Figure 4.7: Bubble map for GO terms enrichment analysis of miRNAs. The y-axis identifies the GO term, the x-axis defines the enrichment (-log10(p-value)), -log10(p-value) are represent by the color scale, and the number of genes is represented by the size of the nodes.

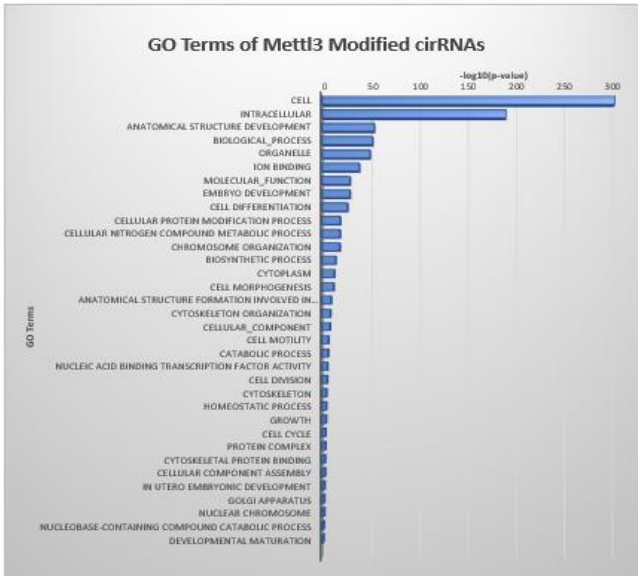

Figure 4.8: GO terms analysis of miRNAs. The y-axis displays the GO term, and the x-axis displays the -log10(p-value).

The generated KEGG and GO analysis will help with understanding the regulatory functions of the Mettl3 modified cirRNAs and how these functions match or differ from their compartment linear mRNAs.

CHAPTER 5

DISCUSSION

In this chapter, the results section will be discussed, and our findings will be presented. This chapter will focus on answering the two important questions presented in the Introduction chapter by comparing the KEGG pathways and GO terms generated by the Mettl3 modification of mRNAs and cirRNAs.

5.1 Does Mettl3 regulates the same biological pathways and processes in both linear mRNAs and cirRNAs?

To answer this question, the biological pathways and processes that get regulated by Mettl3 for both the linear mRNAs and cirRNAs will be compared. The research study of "Distinct roles of Fto and Mettl3 in controlling the development of the cerebral cortex through transcriptional and translational regulations" that was mentioned earlier and used as the starting point of our research, highlighted the different biological pathways and processes of the linear mRNAs that get regulated by Mettl3. It was found by scientists that the modifications of mRNAs resulted in the regulation of many KEGG biological pathways affecting the mice's embryonic cerebral cortex development. Of these biological pathways, signal transduction, signaling molecules and interactions, degradation, cell growth and death, transport and catabolism, protein folding and translation, and axonal fasciculation were mentioned. It has been also confirmed that Mettl3 can initiate and develop the progression of different types of cancers as suggested by many earlier studies. Also, the

GO analysis of the study presented the different biological processes regulated by the modification of linear mRNAs, and these processes include cell differentiation, neurogenesis (the process of generating cells within the nervous system), neuron differentiation (enhancing the cells with specialized neurons), cell cycle, regulation of cell prefoliation, and cell cycle process. Additionally, the alteration of mRNAs by knocking out Mettl3 led to an alteration in the translation of the cortical glial cells' genes affecting the development of the central nervous system.

Comparing the KEGG pathways and GO categories obtained for the cirRNAs (from our study) and the linear mRNAs (from the reference article as discussed above) indicate that all of the KEGG biological pathways that are regulated by the Mettl3 modification of mRNAs have also been regulated by the corresponding cirRNAs. These cirRNAs regulate the gene expression of multiple biological pathways by being rich in miRNAs binding sites and work as sponges to control the functions of miRNAs.

However, two GO terms appear to be mentioned in the research paper but were not established for our cirRNAs. These two biological terms are the neurogenesis and neuron differentiation terms. These two biological terms/processes are highly important for the development of brains in embryos. This indicates that not all the gene regulatory functions that are controlled by the Mettl3 modified mRNA are also controlled by the Mettl3 modified cirRNAs. This finding shows that Mettl3 modified linear mRNAs might be specialized in regulating some unique biological processes that cannot be regulated by the corresponding Mettl3 modified cirRNAs which could be contributed to the different miRNA binding affinities for both molecules.

In summary, Mettl3 modified cirRNAs appear to participate in gene regulation and controlling biological pathways and processes in a manner that is very comparable but not exactly the same as their corresponding linear mRNAs.

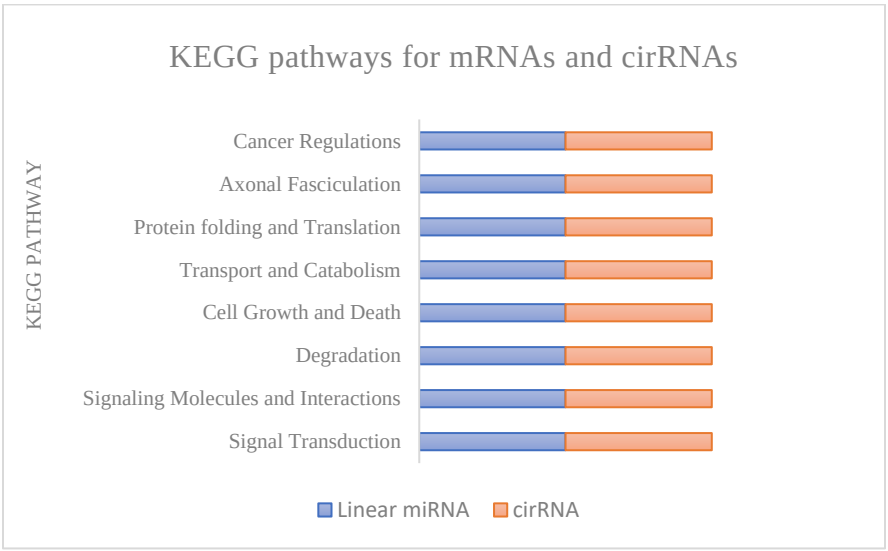

Figure 5.1: The KEGG pathways of both the Mettl3 modified linear mRNAs and cirRNAs.

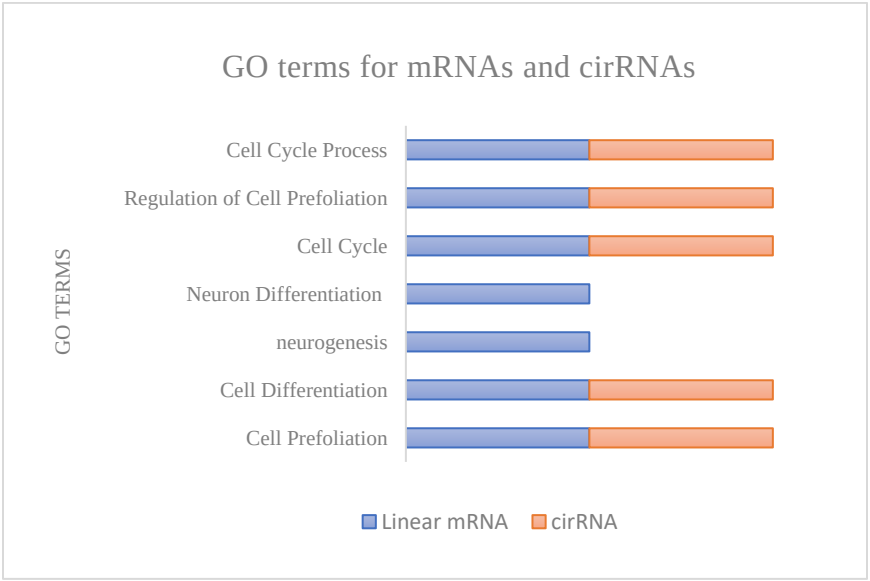

Figure 5.2: The GO terms of both the Mettl3 modified linear mRNAs and cirRNAs.

5.2 Are there any unique biological pathways and processes initiated by the Mettl3 modification in cirRNAs that are not being detected in linear mRNAs?

In addition to the biological pathways that were regulated in both linear mRNAs and cirRNAs, interestingly, additional biological pathways and processes have been regulated by the Mettl3 modified cirRNAs and were not regulated in the linear mRNAs. The first important biological KEGG pathways that get regulated in cirRNAs is the fatty acid metabolism and biosynthesis. This finding shows that cirRNAs have the potential of regulating the fatty acid synthesis which is essential for developing the mice's embryonic cerebral cortex that controls different brains activities including consciousness, emotion, reasoning, and memory. The second KEGG pathway that was regulated in cirRNAs is the endocytosis pathway. In neuron cells, endocytosis is the process that the cells take to absorb the nutrients and other components of the plasma membrane. This process is critical for the early stages of development and any modifications to this pathway can result in many neurological diseases such as brain tumors and intellectual disabilities. Additionally, cytoskeleton and its cytoskeletal protein binding GO processes also appeared to be regulated in cirRNAs but not in linear mRNAs. The Cytoskeleton is a structure that is composed of protein filaments and functions as mechanical support that maintains the shape of the cells. The actin filaments in the cytoskeleton play an important role in the regulation of the cellular activities in the brain. An additional pathway that appears to be regulated in the cirRNAs is the Golgi apparatus. Golgi apparatus is a cell structure that works on processing and packaging lipids and proteins. In the neural system, the Golgi apparatus plays an essential role in regulating the pathways of the ion channels, the signaling molecules, and the receptors which affect embryonic neuronal development.

These additional biological pathways and processes that were regulated by the Mettl3 modifications of cirRNAs introduce the biological and functional importance of cirRNAs that can be involved in the regulation of gene expression in a manner that differs from their corresponding linear mRNAs. These important biological pathways and processes lead us to the conclusion that Mettl3 modified cirRNAs have the potential of regulating embryonic cerebral cortex development in mice by controlling some additional pathways that were not regulated by Mettl3 modified mRNAs. These pathways include fatty acid metabolism and biosynthesis, endocytosis pathway, cytoskeleton, and its cytoskeletal protein binding processes, and Golgi apparatus which all play a critical role in the embryonic cerebral cortex neuronal development in mice.

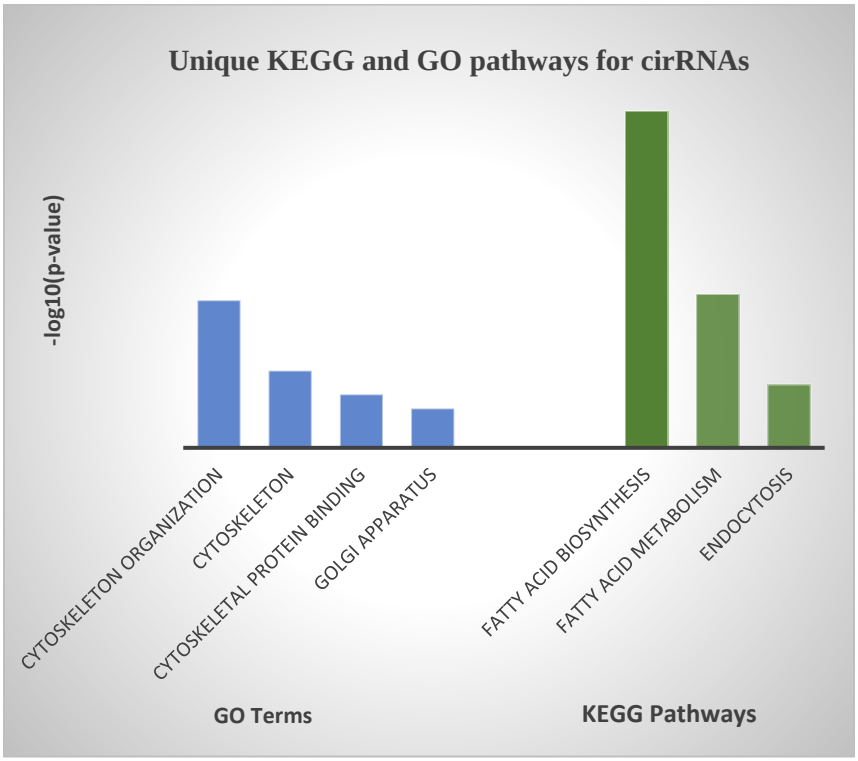

Figure 5.3: The unique KEGG pathways and GO terms generated for the Mettl3 modified cirRNAs.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, the role of Mettl3 modification of cirRNAs in regulating gene expression and controlling different KEGG biological pathways and GO processes in the embryonic cerebral cortex of mice was studied. Although many studies have been conducted to understand the regulatory roles of Mettl3 modified mRNAs, fewer contributions were applied to understand the roles of Mettl3 modified cirRNAs on gene expression and on the regulation of different biological pathways and processes. cirRNAs have been proven to be associated with the occurrence and progression of many diseases including different types of cancers, autoimmune diseases, neurological disorders, infertility disorders, and diabetes. Still, a limited number of studies have been applied to understand their regulatory contributions to different KEGG biological pathways and GO terms. Accordingly, the roles of Mettl3 modified cirRNAs in the embryonic cerebral cortex development in mice were studied in this research by finding the cirRNAs that are significantly enriched in miRNA binding motifs. Having multiple miRNA binding sites give the cirRNAs their "miRNA sponge" function which is highly significant in regulating gene expression and biological pathways at both transcriptional and translational levels.

It has been concluded that Mettl3 modified cirRNAs can regulate gene expression and control different biological pathways and processes in a manner that is similar but not identical to their corresponding linear mRNAs. Both Mettl3 modified mRNAs and cirRNAs regulate the following KEGG biological pathways: transport and catabolism,

cancer regulation, axonal fasciculation, protein folding and translation, cell growth and death, degradation, signaling molecules and interactions, and single transduction. GO terms, on the other hand, show that some additional processes were regulated in mRNAs but not in cirRNAs. These GO terms are neurogenesis and neuron differentiation which is both critical to the development of brain functions in embryos. Interestingly, it has been found that Mettl3 modification in cirRNAs can promote the regulation of unique biological pathways and processes that are significant to the embryonic cerebral cortex neuronal development in mice. These pathways include fatty acid metabolism and biosynthesis, endocytosis pathway, cytoskeleton, and its cytoskeletal protein binding processes, and Golgi apparatus that are all critical in the development of different cellular activities in the brain and the regulation of the progress of neurological diseases such as brain tumors and intellectual disabilities in the embryonic cerebral cortex of mice.

The developed framework that was followed in this study can be generalized for any future studies focusing on understanding the gene regulatory roles and biological contributions of cirRNAs in different organisms and within different biological tissues. The finding of the Mettl3 regulatory roles of cirRNAs in the embryonic cerebral cortex of mice as established by this study will be essential for the future studies that aim to use cirRNAs as diagnostic biomarkers and therapeutic targets by understanding their role in disease progression and due to their stable, abundant, and conservative nature within different biological tissues.

Future studies should also focus on identifying the biological pathways regulated by cirRNAs that appear within the differently expressed conditions. The differently expressed conditions include the existence of cirRNAs in both normal and disease conditions but are more abundant in one of these conditions. This will help with the

identification of the regulatory roles and the biological pathways controlled by the cirRNAs relative to their abundance.

REFERENCES

[1] Balakrishnan, Rama, et al. “A Guide to Best Practices for Gene Ontology (GO) Manual Annotation.” *Database: the Journal of Biological Databases and Curation*, Oxford University Press, 9 July 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706743/.

[2] Chaabane, Mohamed, et al. “Seekcrit: Detecting and Characterizing Differentially Expressed Circular RNAS Using High-Throughput Sequencing Data.” *PLOS Computational Biology*, Public Library of Science, https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1008338

[3] “Codon Charts - Codon Table Sheets.” *Genomenon*, 24 Aug. 2021, https://www.genomenon.com/codon-chart/.

[4] Du, Kunzhao, et al. “Distinct Roles of FTO and METTL3 in Controlling Development of the Cerebral Cortex through Transcriptional and Translational Regulations.” *Nature News*, Nature Publishing Group, 14 July 2021, https://www.nature.com/articles/s41419-021-03992-2#Sec2.

[5] Freeman, Jenny V. and Michael J Campbell. “THE ANALYSIS OF CATEGORICAL DATA: FISHER’S EXACT TEST.” (2007).

[6] Giani, Alice Maria, et al. “Long Walk to Genomics: History and Current Approaches to Genome Sequencing and Assembly.” *Computational and Structural Biotechnology Journal*, Elsevier, 17 Nov. 2019, https://www.sciencedirect.com/science/article/pii/S2001037019303277.

[7] GLICK, BERNARD R. “Fundamental Technologies.” *Molecular Biotechnology: Principles and Applications of Recombinant DNA*, 5th ed., AMER SOC FOR MICROBIOLOGY, S.l., 2022, pp. 11–91.

[8] Hammond, Scott M. “An Overview of Microrna.” *Advanced Drug Delivery Reviews*, U.S. National Library of Medicine, 29 June 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4504744/.

[9] Heather, James M, and Benjamin Chain. “The Sequence of Sequencers: The History of Sequencing DNA.” *Genomics*, Academic Press, Jan. 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4727787/#bb0075.

[10] Hunter, and m. biology, “Molecular biology for computer scientists,” pp. 1-46, 1993.

[11] Hrdlickova, Radmila, et al. “RNA-Seq Methods for Transcriptome Analysis.” *Wiley Interdisciplinary Reviews. RNA*, U.S. National Library of Medicine, Jan. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5717752/.

[12] Kukurba, Kimberly R, and Stephen B Montgomery. “RNA Sequencing and Analysis.” *Cold Spring Harbor Protocols*, U.S. National Library of Medicine, 13 Apr. 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/.

[13] Kozomara, Ana, et al. “MiRBase: from MicroRNA Sequences to Function.” *Academic.oup.com*, 8 Jan. 2019, https://academic.oup.com/nar/article/47/D1/D155/5179337.

[14] Laboratories, Kanehisa. “Genomics to Biological System.” *Kegg Overview*, https://www.genome.jp/kegg/kegg1a.html#:~:text=Genomes%20to%20Biological%20System,genomic%20and%20molecular%2Dlevel%20information.

[15] Macfarlane, Leigh-Ann, and Paul R Murphy. “MicroRNA: Biogenesis, Function and Role in Cancer.” *Current Genomics*, Bentham Science Publishers Ltd, Nov. 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048316/#:~:text=MicroRNA%20(miRNA)%2C%20originally%20discovered,genes%20%5B4%2D8%5D.

[16] O'Brien, Jacob, et al. “Overview of Microrna Biogenesis, Mechanisms of Actions, and Circulation.” *Frontiers*, Frontiers, 1 Jan. 1AD, https://www.frontiersin.org/articles/10.3389/fendo.2018.00402/full.

[17] Ogata, Hiroyuki, et al. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Academic.oup.com*, https://academic.oup.com/nar/article/27/1/29/1238108.

[18] Ren, Xiaoxia, et al. “Potential Functions and Implications of Circular RNA in Gastrointestinal Cancer (Review).” *Oncology Letters*, Spandidos Publications, 1 Dec. 2017, https://www.spandidos-publications.com/10.3892/ol.2017.7118.

[19] Sigrid Rouam, and Sigrid Rouam1Email author. “False Discovery Rate (FDR).” *SpringerLink*, Springer, New York, NY, 1 Jan. 1970, https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_223.

[20] Szabo, Linda, et al. “Statistically Based Splicing Detection Reveals Neural Enrichment and Tissue-Specific Induction of Circular RNA during Human Fetal Development.” *Genome Biology*, BioMed Central, 16 June 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4506483/.

[21] UofLBioinformatics. “UofLBioinformatics/Seekcrit: Seek for Circular RNA in Transcriptome (Tool to Identify Deferentially Expressed Circrnas between Two Samples).” GitHub, https://github.com/UofLBioinformatics/seekCRIT.

[22] Wang, Zhong, et al. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews. Genetics*, U.S. National Library of Medicine, Jan. 2009, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/.

[23] Zeng, Xiangxiang, et al. “A Comprehensive Overview and Evaluation of Circular RNA Detection Tools.” *PLOS Computational Biology*, Public Library of Science, https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1005420

[24] Zhang, Xiao-Ou, et al. “Diverse Alternative Back-Splicing and Alternative Splicing Landscape of Circular RNAS.” *Genome Research*, Cold Spring Harbor Laboratory Press, Sept. 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5052039/.

CURRICULUM VITAE

NAME: Dunia Zedan

ADDRESS: Computer Engineering & Computer Science Department

University of Louisville

Louisville, KY 40292

EDUCATION:

Master of Science: Computer Science
Expected Graduation: May 2022
University of Louisville – Louisville, KY

Bachelor of Science: Chemistry/Biochemistry, Fall 2017
University of Louisville – Louisville, KY

CERTIFICATE:

Graduate Certificate in Data Science, Fall 2021
University of Louisville – Louisville, KY

HONORS AND AWARDS:

- CSE Master of Science, April 2022
- Magna Cum Laude, Fall 2017
- University of Louisville Dean's Honors, Spring 2016 – Fall 2017
- Golden Key International Honor Society

SCHOLORSHIP:

- Academic Commonwealth Scholarship (5,300), Fall 2016 – Fall 2017

RESEARCH EXIPERANCE:

- University of Louisville, Department of Computer Engineering & Computer Science, Spring 2022
  Advisor: Dr. Juw Won Park
  Discovering the pathways and GO terms associated with Mettl3 modified circular RNAs in the embryonic cerebral cortex of mice.

- University of Louisville, Department of Chemistry, Fall 2017
  Advisor: Dr. Cecilia Yappert
  Using Fluorescent Rhodamine Dyes to Detect Phospholipids and Steroids with MALDI-TOF Mass Spectrometry.