Circular RNA Characterization and Regulatory Network Prediction in Human Tissue

by

Shobana Sekar

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2018 by the
Graduate Supervisory Committee:

Winnie S Liang, Co-Chair
Valentin Dinu, Co-Chair
David W Craig
Li Liu

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

Circular RNAs (circRNAs) are a class of endogenous, non-coding RNAs that are formed when exons back-splice to each other and represent a new area of transcriptomics research. Numerous RNA sequencing (RNAseq) studies since 2012 have revealed that circRNAs are pervasively expressed in eukaryotes, especially in the mammalian brain. While their functional role and impact remains to be clarified, circRNAs have been found to regulate micro-RNAs (miRNAs) as well as parental gene transcription and may thus have key roles in transcriptional regulation. Although circRNAs have continued to gain attention, our understanding of their expression in a cell-, tissue- , and brain region-specific context remains limited. Further, computational algorithms produce varied results in terms of what circRNAs are detected. This thesis aims to advance current knowledge of circRNA expression in a region specific context focusing on the human brain, as well as address computational challenges.

The overarching goal of my research unfolds over three aims: (i) evaluating circRNAs and their predicted impact on transcriptional regulatory networks in cell-specific RNAseq data; (ii) developing a novel solution for *de novo* detection of full length circRNAs as well as *in silico* validation of selected circRNA junctions using assembly; and (iii) application of these assembly based detection and validation workflows, and integrating existing tools, to systematically identify and characterize circRNAs in functionally distinct human brain regions. To this end, I have developed novel bioinformatics workflows that are applicable to non-polyA selected RNAseq datasets and can be used to characterize circRNA expression across various sample types and diseases. Further, I establish a reference dataset of circRNA expression profiles and

regulatory networks in a brain region-specific manner. This resource along with existing databases such as circBase will be invaluable in advancing circRNA research as well as improving our understanding of their role in transcriptional regulation and various neurological conditions.

I would like to dedicate this to my parents who always showered me with their unconditional love, support and encouragement through my success and failures, and my dearest late grandfather, Mr. Krishna Iyer.

ACKNOWLEDGMENTS

I would like to thank my entire supervisory committee for all their guidance and support throughout my project. I am deeply indebted to my mentor and co-chair, Dr. Winnie Liang, without whom I would not have reached this far. I would like to thank Dr. Valentin Dinu, Dr. David Craig, Dr. Li Liu and Dr. Garrick Wallstrom for all their valuable insights and guidance throughout my research. I would also like to express my deepest gratitude to Dr. Jonathan Keats and Dr. Sara Nasser, TGen for sharing their expertise and helping me with my thesis. I am also grateful to Dr. Kendall Jensen, Dr. Elizabeth Hutchins and Jessica Aldrich, TGen for all their valuable guidance and suggestions. Last but far from least, I would like to thank my loving father, mother, brother and husband, for all their love and encouragement through every step of the way.

Research reported in this thesis was supported by the National Institute on Aging (NIA) of the National Institutes of Health under award number P30AG019610, and the Arizona Department of Health Services award number ADHS14-052688. The content is solely the responsibility of the author and the funders had no role in the study design, data collection and analysis, decision to publish, or preparation of manuscripts.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ACValidator - Assembly based circRNA Validator

AD - Alzheimer's Disease

ADAR1 - Adenosine Deaminase 1

AGO – Argonaute

ALDH1L1 - Aldehyde Dehydrogenase 1 family, member L1

A-to-I - Adenosine-to-Inosine

BAM - Binary Alignment Mapping

BBDP - Brain and Body Donation Program

BC - Brain Cerebellum

BCL - Base Call file

BSHRI - Banner Sun Health Research Institute

cANRIL - circular ANRIL

CDS - Coding DNA Sequence

CIGAR - Concise Idiosyncratic Gapped Alignment Report

CircRNA - Circular RNA

CIRI - Circular RNA Identifier

CiRS-7 - circRNA Sponge for miR-7

CSCD - Cancer Specific circRNA Database

dbGaP - database of Genotypes and Phenotypes

DCC - Detect circRNAs from Chimeric reads

DE - Differentially Expressed

DeFuCir - Detection of Full length circRNAs

EIcircRNA - Exon-Intron circRNAs

FDR - False Discovery Rate

GBM - Glioblastoma Multiforme

GLM - Generalized Linear Model

IGV - Integrated Genomics Viewer

IPL - Inferior Parietal Lobe

IRES - Internal Ribosomal Entry Site

KEGG - Kyoto Encyclopedia of Genes and Genomes

lncRNA - long non-coding RNA

LOAD – Late-Onset AD

MG - Middle Temporal Gyrus

miRNA - microRNA

mRNA - messenger RNA

NCBI - National Center for Biotechnology Information

ND - No Disease

NIA - National Institute on Aging

OC - Occipital Cortex

ORF - Open Reading Frame

PAR-CLIP - Photo Activatable-Ribonucleoside-enhanced Cross Linking and Immunoprecipitation

PC - Posterior Cingulate

PCR - Polymerase Chain Reaction

PKA - Protein Kinase A

PSAP – Prosaposin

PTM - Post-Translational Modification

QKI – Quaking

RBP - RNA-Binding Proteins

RNase R - RiboNuclease R

RNAseq - RNA sequencing

rRNA - ribosomal RNA

SAM - Sequence Alignment Mapping

SF - Superior Frontal Gyrus

siRNA - small interfering RNA

SNV - Single Nucleotide Variants

SRA - Sequence Read Archive

tricRNA - tRNA intronic circRNA

tRNA - transfer RNA

TSCD - Tissue Specific circRNA Database

UTR - Untranslated Region

CHAPTER 1

INTRODUCTION

1. Background

The transcriptome of an organism refers to the set of all RNA molecules in a cell or a population of cells. Traditionally, the whole transcriptome was considered to be primarily composed of messenger RNAs (mRNAs), ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). However, more recently additional RNA species such as microRNAs (miRNA), piwiRNAs, and long non-coding RNAs (lncRNAs) have been characterized. Circular RNAs (circRNAs) represent one such recent addition to the whole transcriptome. CircRNAs are endogenous non-coding RNAs that form covalently closed continuous loops and are both highly conserved and abundant in the mammalian brain [1-3]. Though circularization events in the RNA were reported in the 1970s and 1990s [4-6], they were disregarded as molecular artifacts formed as a result of aberrant splicing. However, with the advent of next generation sequencing technology coupled with the development of computational algorithms focused specifically on detecting these back-splicing events, numerous circRNAs have been reported since 2012.

Some of the key findings from these circRNA studies include (i) circRNAs exhibit cell type, tissue and developmental stage specific expression [3, 7]; (ii) they have no 5'-3' polarity or polyadenylated tail, making them more stable than linear RNAs [2]; (iii) they exhibit evolutionary conservation between the mouse and human [2, 3]; (iv) they tend to have longer flanking introns (compared to non-circularized expressed exons) that are enriched with reverse complementary sequences, possibly aiding in their

biogenesis [2]; (v) in some cases, they exhibit higher expression than their cognate linear isoforms [1-3]; (vi) most exonic circRNAs are localized to the cytoplasm [1-3]; (vii) naturally occurring circRNAs are not associated with ribosome protected fragments, suggesting that they are not translated; however, engineered circRNAs designed with an internal ribosomal entry site (IRES) can be translated *in vitro* and *in vivo* [8, 9].

2. Biogenesis

The majority of identified circRNAs are derived from exons, wherein a downstream splice donor is covalently linked to an upstream splice acceptor, to form a 'head-to-tail' splice junction.

Although the exact mechanism of circRNA biogenesis remains to be elucidated, a few possible mechanisms have been proposed:

2.1 Lariat driven circularization

When the middle exons are skipped during a linear splicing event, the spliced out intron lariat contains those skipped exons. If further splicing occurs within the lariat before its degradation, a stable circRNA enclosing the skipped exons may be generated. 45% of the predicted circRNAs in [2] also exhibited exon skipping events, suggesting that RNA circularization could be correlated with exon skipping.

2.2 Intron-pairing driven circularization

Intronic motifs that flank the potential circRNA exons such as ALU repeats, complementary sequences or other RNA secondary structures could also aid in exon

circularization. In this model, exon skipping is not required since the complementarity of the flanking sequences brings the non-sequential donor acceptor pairs side by side [2].

Apart from these, specific RNA-binding proteins (RBPs) have been reported to regulate circRNA biogenesis, such as quaking (QKI) and adenosine deaminase 1 (ADAR1). QKI regulates circRNA formation during epithelial-mesenchymal transition in humans, and the insertion of synthetic QKI-binding sites into linear RNA induces exon circularization [10]. ADAR1 on the other hand suppresses specific circRNA expression by the adenosine-to-inosine (A-to-I) editing activity of ADAR1 on the duplexes formed between the circRNA flanking sequences. Furthermore, knockdown of ADAR1 significantly up-regulates the expression of specific circRNAs, suggesting that ADAR1 antagonizes circRNA expression [11].

Although the majority of reported circRNAs are derived from exons, special subtypes of circRNAs have also been observed. These include exon-intron circRNAs (EIcircRNAs), composed of both exonic and intronic sequences [12], circular intronic RNAs formed from eukaryotic spliceosomal introns [13], and tRNA intronic circRNA (tricRNAs) generated during pre-tRNA splicing [14].

3. Putative functions

The abundance and evolutionary conservation of circRNAs suggests that they could have a potential role in cellular processes, the majority of which are still unknown. However, a few possible functions have been reported, including micro-RNA (miRNA) binding, mediation of protein-protein interactions and regulation of parental gene transcription [3, 6, 12, 15-19].

3.1 MicroRNA regulation

MicroRNAs are 18 to 25 nucleotides long and are a class of non-coding RNAs. They are important regulators of gene expression and function by directly pairing to 3' untranslated regions (UTRs) of their target mRNAs. Two exonic circRNAs, one from the mouse *SRY* gene [6, 15] and the other, human/mouse cirRS-7 (also known as *CDR1as)* [3, 16] have been shown to bind miRNAs. Further, photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) experiments have been used to validate CDR1as binding to miR-7 effector protein argonaute (AGO) [3]. The circular *Sry* transcript harbors 16 miR-138 binding sites, while cirRS-7 has 74 miR-7 binding sites. Notably, the cirRS-7 circRNA is highly expressed in human and mouse brain and acts as a sponge/inhibitor of miR-7, thus resulting in increased levels of miR-7 targets [3]. Such miRNA binding competes with the binding of miRNAs to their linear transcript targets, thereby mitigating the effect of miRNA-mediated post-transcriptional regulation.

3.2 Protein decoys

CircRNAs could also act as decoys to RBPs, thereby impacting functions in which RBPs are involved. Recently, circRNA circ-Foxo-3 was found to form a ternary complex with cell cycle proteins cyclin-dependent kinase 2 (CDK2) and cyclin-dependent kinase inhibitor 1 (p21). This interaction prevents CDK2 from interacting with cyclins A and E, which is required for cell cycle progression [19]. In addition, circ-Foxo3 interacts with senescence response associated proteins such as ID1, E2F1, FAK and HIF1A and promotes cardiac senescence [18].

3.3 Transcriptional and splicing regulation

Some circRNAs have also shown the ability to regulate parental/host gene transcription. In a recent study, the EIcircRNA EIciEIF3J was found to interact with the U1 spliceosomal component and the promoter of *EIF3J* to enhance *EIF3J* transcription [12]. In a separate study, the interaction between the circular form of musclebind circMbl and MBL protein was found to modulate the splicing activity of MBL and regulate the pre-mRNA splicing of the host gene by competing with the canonical splicing machinery [17].

4. CircRNAs in brain and other tissue types

Recent studies have reported preferential back-splicing of neural genes and an abundance of circRNAs in the mammalian brain. Rybak-Wolf *et al*. [20] analyzed human and mouse neuronal cell line data as well as published sequencing data from the ENCODE Consortium [21] to identify circRNAs. Their analysis revealed that circRNAs are highly abundant in the mammalian brain compared to other analyzed tissues such as lungs, heart, kidney, testis and spleen, with well conserved sequences across mouse and humans. They also found that circRNAs were up-regulated during neuronal differentiation and development, and highly enriched in synapses, independent of their corresponding linear isoform.

In a separate study [22], analysis of RNA sequencing (RNAseq) data from mouse brain, liver, heart, lung and testis revealed that although circRNAs were present in all the examined tissues, their abundance was highest in the brain. Further, 20% of the protein coding genes in the brain were found to produce circRNAs. The study also found an

enrichment of circRNAs compared to their host linear transcripts in synaptosomes and microdissected neuropils from mouse hippocampal slices. Analysis of circRNA expression in developing cultured hippocampal neurons (stages: E18 - embryonic, P1 - early postnatal, P10 - postnatal at the beginning of synapse formation, and P30 - late postnatal) revealed an increase in circRNA levels during the P10 - P30 transition, corresponding to the time of synapse formation. Notably, the up-regulated circRNAs during this transition were mostly derived from host mRNAs that play a role in synaptic function such as *Dlgap1* (DLG associated protein 1) and *Homer1 (homer scaffolding protein 1).*

Various other tissue types including heart, colon, kidney, liver and lung, as well as glandular tissues such as the adrenal gland, mammary gland, pancreas and thyroid gland have also been examined and found to have varying abundances of circRNAs (Table 1.1) [23-26]. Apart from these, extracellular fluids such as plasma and serum have also been investigated [25]. The number of circRNAs detected from these studies has been summarized in Table 1.1.

Study	Tissue type examined	Number of Donors	Number of circRNAs detected	Ref
Tan et al., 2016	Heart	12	15318	26
Xu et al., 2016	Colon, heart, kidney, liver, stomach	6	8,120	28
Xu et al., 2016	Adrenal gland, mammary gland, pancreas, thyroid gland	4	14,433	27
Maass et al., 2017	Placenta	1	63	
Maass et al., 2017	Umbilical cord	1	85	
Memczak et al., 2015	Cerebellum	2	6,289	25
Memczak et al., 2015	Liver	2	1,198	
Memczak et al., 2015	Blood	2	6,213	
Maass et al., 2017	Plasma	1	57	27
Maass et al., 2017	Serum	1	39	

Table 1.1: Summary of circRNA detection results in various tissue types

5. CircRNAs in other species

Apart from humans, circRNAs have also been observed in several other species, particularly mice [1-3, 20, 22, 27-29], across various studies. Some of the initially identified circRNAs were from the mouse *SRY* (sex determining region-Y) gene [6] and the *DCC* (Deleted in Colorectal Carcinoma) gene in human and rodent cells [5]. Other common model organisms used in circRNA studies include *Drosophila melanogaster* [7, 30] and *Caenorhabditis elegans* (C. elegans) [3, 31]. Memczack *et al.* [3] were one of the first few to report circRNAs in other species, and identified 1,903 circRNAs in mouse sequencing data and 724 circRNAs in *C. elegans*, each with at least two independent junction spanning reads [3]. They also observed for the first time that in the nematode, some of the expressed circRNAs showed developmental stage specific expression.

6. CircRNAs in disease

Recent evidence has suggested that circRNAs may have regulatory roles in the initiation and development of diseases.

6.1 Alzheimer's disease

Given the high abundance of the miRNA miR-7 and its association with ciRS-7 in the human brain, Lukiw [32] studied the levels of ciRS-7 in Alzheimer's affected brain cells from the CA1 hippocampal region. Using Northern blots and ribonuclease R (RNase R) treatment, his group observed that ciRS-7 levels were significantly reduced in AD brains, suggesting the existence of a dysregulated ciRS-7-miR-7 system in these samples. RNase R is a magnesium-dependent 3´-5´ exoribonuclease (enzyme that degrades RNA from the terminal nucleotides) that digests most linear RNAs but leaves behind lariat or circRNA structures. The study hypothesizes that due to the low abundance of ciRS-7, less regulation of miR-7 may down-regulate expression of AD associated targets such as ubiquitin protein ligase A, which is normally responsible for clearing up amyloid plaques in AD. However, functional studies still need to be performed to elucidate the role of ciRS-7 in AD.

6.2 Glioblastoma

In order to explore the expression profiles of circRNAs in glioblastoma multiforme (GBM), Zhu *et al.* [33] analyzed tumor tissues from five GBM patients and five normal brain samples. Their study revealed a total of 1,411 differentially expressed circRNAs in GBM patients, among which 206 were up-regulated and 1,205 were down-regulated circRNAs. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis on the down-regulated circRNAs revealed that ErbB signaling pathway and

Neurotrophin signaling pathway were the most enriched. Further, circBRAF expression was found to be significantly higher in normal brain tissues compared to glioma tissues. Additionally, circBRAF was significantly lower in glioma patients with a higher pathological grade than those with lower grades.

6.3 Acute lymphoblastic leukemia

Salzman and colleagues hypothesized that local structural rearrangements in the human genome may contribute to the progression of human cancers [1]. In order to test this hypothesis, the authors performed RNAseq on rRNA depleted total RNA from the bone marrow of 5 children with acute lymphoblastic leukemia (ALL). They searched for transcripts with exons in a scrambled order. This analysis identified hundreds of genes with exon scrambling and these factors were estimated to comprise more than 10% of total transcript isoforms. Contrary to their hypothesis, they found that all the PCR-verified exon scrambling events in leukemia samples were also present in HeLa cells and peripheral blood cells collected from the same patients, as well as in H9 human embryonic stem cells. This seemed to suggest that the splicing process that led to the formation of scrambled exons were active in both normal and malignant human cell types. Of the two models considered to explain transcripts with scrambled exons, their statistical approach predicted that scrambled exons originated in a circRNA. To experimentally validate this finding, a panel of 9 isoforms across 6 genes with scrambled exons was tested for sensitivity to RNase R. The resistance of these transcripts to RNase R degradation supported their hypothesis that the majority of the scrambled transcripts were circRNAs.

6.4 Colorectal cancer

Heyda et al. [34] looked for circRNAs in rRNA depleted RNAseq datasets from 12 tumor tissues of colorectal cancer patients and their matched normal colon mucosa samples. Their analysis revealed an over-representation of significantly down-regulated circRNAs in the tumor samples. Further, they found that the ratio of circular to linear RNA isoforms was lower in tumor samples compared to normal colon samples and that this ratio correlated negatively with the proliferation index (percentage of Ki-67 positive cells; Ki-67 is a protein required for cellular proliferation). The negative correlation of global circRNA abundance to proliferation was validated in rRNA depleted RNAseq datasets from idiopathic pulmonary fibrosis samples, serous ovarian cancer cell samples and 13 normal human tissues. In all three, the expression level of *MKI67* (Maker of proliferation Ki-67) was negatively correlated to global circRNA abundance.

7. Circular RNA detection

7.1 Biochemical enrichment of circular RNAs

In 2010, Burd et. al [35] discovered a circRNA species, circular ANRIL (cANRIL), expressed at very low levels in Hs68 human fibroblast cell lines, and which correlated positively with the risk of atherosclerosis. Following this discovery, they developed an experimental protocol for circRNA enrichment, CircleSeq, wherein rRNA is first depleted from total RNA using RNase R, and high-throughput sequencing is performed on the RNA library [2]. Using this protocol, they enriched for circRNAs in human fibroblast cells and identified circRNAs using Mapsplice [36], a *de novo* splice

mapping algorithm. Notably, this enrichment step has also been successfully used by other groups to enrich for circRNAs [3, 24].

7.2 Computational tools for circular RNA detection

In the last few years, various circRNA prediction algorithms have been developed in order to computationally detect circRNAs from RNAseq data. These algorithms can be classified into two broad categories based on their dependency on genome annotation: pseudo-reference-based and fragment-based strategies. In the pseudo-reference-based strategy, a pseudo reference comprised of all possible combinations of exon pairs is constructed. RNAseq reads aligning to these putative circRNA references are marked as potential circRNAs. On the other hand, the fragment-based strategy (also known as segmented read based strategy) does not require a genome annotation, but detects circRNAs based on mapping information of a split read's alignment to the reference genome (split read refers to a sequencing read aligning to two distinct portions of the genome). When segments of a split read align to the reference in non-colinear order, they are marked as potential circRNA candidates by the tools. A discussion of the different tools is presented below and briefly summarized in Table 1.2.

Tool	Strategy	Aligner	Filtering rules	Number of circRNAs detected	Blind spots/ drawbacks
Find_ circ	Segmented read based	Bowtie2 [37]	GT-AG splice signals; 100 kb distance; Unique anchor alignment; Anchor extension aligns completely; Break point resides within 2 nucleotides inside the anchor; Unambiguous breakpoint	1,950 (human); 1,903 (mouse); 724 (nematode)	Non-canonical splice signals not considered
CIRI	Splicing signal based	BWA-MEM [38]	GT-AG splice signals; Filter out circRNAs in homologous genes or repeat regions	3001 (human)	Non-canonical splice signal detection requires gene annotation file
Map-splice	Segmented read based	Bowtie [39]	Mismatch rate; base call quality and junction score	7771 (human); 646 (mouse)	Computational time, requires gene annotation
KNIFE	Pseudo-reference based	Bowtie, Bowtie2	Reads aligning to linear junctions removed; User-specified junction overlap; Mate maps within the circle; Unique alignment with <= 2 mismatches for de novo detection; UCSC KnownGene annotated exons within 1 Mb	11,965 (human)	CircRNAs between exons > 100 kb apart not considered
DCC	Segmented read based	STAR [40]	R2 maps within circle; Repetitive or homologous genes GT-AG	2,382 (human)	Non-canonical splice signals not considered; requires gene annotation
CIRC-explorer	Segmented read based	STAR/ Tophat [40, 41]	presence of XF tags; only reads that align on same chromosome included; Uniquely aligned reads	2,119 [poly(A)-]; 9639 [poly(A)-/RNase R+] (human)	Non-canonical splice signals and circRNAs from multiple genes not considered Computational time for Tophat

Table 1.2: Summary of circRNA detection tools

7.2.1 Find_circ

Find_circ [3] is one of the first few circRNA detection algorithms developed and that uses the Bowtie2 [37] aligner. The algorithm first aligns RNAseq reads to the reference genome and extracts unmapped reads from this alignment. Next, 20-mer anchors are extracted from both ends of the unmapped reads and re-aligned to the reference genome to identify where they align in spliced exons. Anchors aligning in the reverse orientation are marked as potential circRNAs. The anchor alignments are further extended so that the complete read aligns and GU-AG splice sites flank the break point. Quality cut-offs as follows are then applied on the resulting list of potential candidates (1) unambiguous breakpoint detection; (2) a maximum of only two mismatches in the extension procedure; (3) the breakpoint cannot reside more than 2 nucleotides inside an anchor; (4) at least two independent reads support the junction; (5) unique anchor alignments; difference between best and next-best alignment scores of both anchors above 35 points; and (6) a genomic distance between the two splice sites of no more than 100 kb. Candidates satisfying these filtering criteria are then collated to produce the final catalog of circRNAs.

7.2.2 Mapsplice

Mapsplice [36] is an aligner that can identify multiple types of splice junction events. It is a *de novo* splice mapping tool that can segment reads into multiple anchors to detect canonical and non-canonical (GT-AG and non-GT-AG) splice junctions in RNAseq data. Mapsplice is implemented in two phases to identify such alignments. In the first phase or the ‘tag alignment’ phase, candidate alignments of the mRNA tags to the reference genome are determined. Tags with a contiguous alignment fall within an

exon and can be mapped directly to the reference genome, but those that span splice junctions require a gapped alignment. In the second phase known as the 'splice inference phase', splice junctions in the tag alignments are analyzed to determine a splice significance score based on the alignments and to generate the most likely spliced alignment for each tag. For circRNAs, reads from the fusion category are filtered for splice junctions on the same strand and within 2Mb, but in non-linear order.

7.2.3 CIRCexplorer

CIRCexplorer [42] uses a two-step mapping strategy for circRNA detection. In the first step, sequence reads from each sample are mapped to the reference genome using TopHat2 [41]. Unmapped reads from this step are extracted and mapped to the reference genome using TopHat-Fusion [43]. Reads that split and aligned on the same chromosome but in non-colinear ordering, indicated with special XF tags in output BAM files, are extracted as candidate back-spliced junction reads. These candidate reads are then realigned against existing gene annotations to determine the precise positions of donor or acceptor splice sites. Junction reads with shifted alignments against canonical splice sites are adjusted to the correct positions with custom scripts. Further, reads with alignments on different genes or non-canonical splice sites were largely from trans-splicing or PCR errors, and were therefore discarded.

7.2.4 DCC

DCC (detect circRNAs from chimeric reads) [44] utilizes the output from the splice aware mapper STAR [40] to detect circRNAs. DCC first aligns RNAseq reads against the reference genome. The chimeric reads output by STAR are then filtered based on the following criteria to identify potential circRNA candidates: 1) if paired-end reads

are available, mapping of mates must be within the inferred circRNA; 2) when biological replicates are available, filtering by a minimal number of replicates is performed; 3) the presence of a canonical GT-AG splicing signal at the circRNA junction; 4) mitochondrial circRNAs are discarded; 5) candidates from repetitive or homologous regions are also discarded. DCC also estimates host gene expression from mapping data and reports circular to host gene expression ratios.

7.2.5 KNIFE

KNIFE [45] uses a statistical framework to calculate the posterior probability for every back-spliced read collected to predict whether it is a true positive circRNA. The algorithm first aligns RNAseq reads to the genome, rRNA sequences, and transcriptome references (hg19) as well as to custom linear and scrambled junction sequences using bowtie2. Possible back-splice junction reads that also map with high scores to the other references are discarded. The remaining reads are further classified into circRNA or decoy reads based on the mapping information of the mate when paired-end data are available. A read is considered circular if the mate aligns within the genomic region of the presumed circle defined by the junctional exons, or decoy if the mate aligns outside this region. Further, the reads in the linear and decoy categories are used to fit a generalized linear model (GLM). The GLM predicts the probability that each circular read is a class 1 (true positive) event versus a class 2 (false positive) event.

7.2.6 CIRI

Unlike the above described circRNA detection algorithms, Circular RNA Identifier (CIRI) [46] identifies circRNAs based on the mapping information of the reads, specifically from CIGAR (concise idiosyncratic gapped alignment report) signatures. The

CIGAR string in an alignment file is a sequence of base lengths and associated operation of those bases in alignment, such as match (M), insertion (I), deletion (D) etc. For e.g., 30S115M indicates that in the sequence alignment procedure, 30 bases are soft-clipped and 115 bases match. CIRI takes as input sequence alignment (SAM) files generated with BWA-MEM [38] and scans for segments of a read that align to the reference genome in a chiastic (non-linear) order. To reduce the false positive rate, paired-end mapping and GT-AG splicing signals are used as filtering criteria. The tool then scans the SAM file a second time to identify unbalanced junction reads that were not detected in the first scan, and performs additional filtering to remove candidates resulting from incorrectly mapped reads from homologous genes or repetitive sequences. In addition to exonic circRNAs, CIRI is able to identify circRNAs from intronic and intergenic regions when run on RNAseq datasets from the ENCODE project [21].

Although these circRNA detection algorithms are widely used, one main caveat with bioinformatics based circRNA identification is the highly divergent results produced by the different tools [47, 48]. The use of different aligners, heuristics and filtering criteria introduces variability in the circRNA catalogs produced by the different tools. Further, systematic ‘blind spots’ (false negatives) are also introduced by each tool due to the set of filtering criteria applied [49]. For example, most tools rely on the presence of canonical GT-AG splice signals and thus do not capture candidates with non-canonical splice signals. Most tools use a read count filter, which may not be ideal for circRNAs with low expression relative to their linear host. Given the low reliability on read counts, statistical approaches improve detection and classification of splice junctions, including novel ones. Nonetheless, sequencing errors and technical artifacts introduced during

RNAseq may still lead to false positive circRNAs, and hence statistical tests to estimate false discovery rates in circRNA detection need to be developed.

8. CircRNA databases

Several circRNA databases that have been curated using bioinformatics algorithms are briefly described below.

CircBase [50] is one of the first publicly accessible circRNA databases and enables users to download merged and unified datasets of circRNAs reported from other laboratories. The database is freely accessible through a web server user interface implemented in HTML, CSS and JavaScript and utilizing a MySQL database. Users can query circBase using i) a simple search using identifiers, genomic co-ordinates, sequence etc; ii) a list search that allows the intersection of multiple search terms; iii) a table search that allows for conditional data retrieval. Data from published circRNA studies are added to circBase upon user request and the latest update was performed in July 2017.

Circ2Traits [51] is a knowledgebase of potential associations of circRNAs with human diseases. Using data from Memczack et al. [23], circRNAs associated with disease-related miRNAs were identified, following which the likelihood of circRNA association with a disease is calculated. A network of predicted interactions is constructed between miRNAs, long non-coding RNAs, and circRNAs. The web interface and database were designed using PHP and MySQL, and the interaction networks were built using custom Java programs.

CircNet [52] is a database of tissue-specific circRNA expression profiles and circRNA-miRNA gene regulatory networks. CircNet is based on published circRNA studies as well as 464 RNAseq samples collected from independent studies across 26 human tissues and 104 disease conditions. CircRNAs were identified in these datasets using the algorithm from Memczack et al. [23]. Further, miRNA binding sites were identified by searching for miRNA target sequences in the circRNA isoforms. Through their web interface, users can search for a gene or miRNA of interest to find its associated gene-miRNA-circRNA regulatory network. CircNet also provides genomic annotation of circRNAs with a genome browser integrated into their web interface and was last updated in August 2016. CircRNABase [53] is a similar database that predicts potential circRNA-miRNA interactions, but is based 108 CLIP-seq datasets generated by 37 independent studies. It is thus able to overlap the predicted miRNA target interactions with high throughput CLIP-seq data.

Cancer specific circRNA database (CSCD) [54] is a compilation of the circRNA detection results from 228 total RNA or polyA(-) RNAseq samples from both cancer and normal cell lines. Using CIRI, find_circ, CIRCexplorer and circRNA_finder, a total of 272,152 cancer specific circRNAs were identified, 950,962 were identified in normal samples only and 170,909 were identified in both tumor and normal samples. Further, the microRNA response element sites and RNA binding protein sites for each circRNA were also predicted in silico. Similarly, tissue specific circRNA database (TSCD) compiles circRNA detection results from ENCODE/GEO datasets of 16 adult human tissues (60 samples), 15 fetal human tissues (29 samples) and 9 mouse tissues (24 samples) [55].

A total of 140,681 human adult, 164,069 human fetal and 15,980 mouse tissue specific circRNAs were identified using CIRI, find_circ and circRNA_finder. Open source web framework based on PHP and JavaScript, as well as MySQL tables were used to construct the CSCD and TSCD databases. Users can search for circRNAs by selecting sample type, sample name, gene symbol or the circRNA coordinates.

CircRNAdb [56] is a database of 32,914 human exonic circRNAs selected from diverse sources. For each circRNA, this database provides details such as genomic information, exon splicing, genome sequence, internal ribosome entry site (IRES), open reading frame (ORF) and references. Users can search the web interface of this database using search terms such as chromosome name, gene symbol, transcript, and other keywords.

CircInteractome [57] is a web tool for mapping RBP and miRNA binding sites on human circRNAs. In addition, circInteractome also enables users to (i) design junction-spanning primers to detect circRNAs of interest, (ii) design small interfering RNAs (siRNAs) for circRNA silencing, and (iii) identify potential internal ribosome entry site (IRES). Further, it offers the ability to visualize the corresponding mRNA as well as the genomic and mature sequences of the circRNA.

9. Summary

In summary, circRNAs are a class of RNAs that are gaining more attention in transcriptome research as representing a potential new factor that may influence transcriptional regulation. They are endogenous, non-coding, and abundantly and pervasively expressed in eukaryotes, especially in the mammalian brain. This thesis aims to advance current knowledge of circRNA expression in a tissue and region specific context with emphasis on the human brain, as well as address computational challenges. The overarching goal of our research unfolds over three aims: (i) evaluating circRNAs and their predicted impact on regulatory networks in cell specific RNAseq data (ii) developing assembly based workflows for de novo detection of full length circRNAs as well as in silico validation of selected circRNA junctions (iii) applying the assembly based detection and validation workflows, as well as existing detection algorithms to systematically identify and characterize circRNAs in functionally distinct human brain regions.

In chapter 2, we investigate circRNAs and their potential impact on regulatory networks using RNAseq data derived from posterior cingulate (PC) astrocytes in elderly individuals. To this end, we run existing circRNA detection algorithms and implement an in silico workflow to predict circRNA-miRNA-mRNA interaction networks in these samples. In chapter 3, we present novel assembly based workflows for *de novo* detection of full length circRNAs as well as *in silico* validation of selected circRNA junctions. These bioinformatics workflows are applicable to any non-polyA selected RNAseq dataset and can be used to characterize circRNAs across various sample types and diseases. In chapter 4, we apply these assembly based workflows and integrate results from existing algorithms to identify circRNAs and their regulatory networks in five functionally distinct cortical regions of healthy aged human brain. The brain regions we studied include cerebellum, inferior parietal lobe, middle temporal gyrus, occipital cortex and superior frontal gyrus. Using these analyses, we establish a reference dataset of circRNA expression profiles and regulatory networks in healthy elderly individuals in a region-specific manner. This resource along with existing databases such as circBase will be invaluable in advancing circRNA research as well as understanding their role in transcriptional regulation and various neurological conditions. Lastly, chapter 5 summarizes the findings, conclusions, limitations as well as future directions of this research.

CHAPTER 2

CIRCULAR RNA EXPRESSION AND REGULATORY NETWORK PREDICTION IN POSTERIOR CINGULATE ASTROCYTES IN ELDERLY SUBJECTS

(*The contents of this chapter has been accepted for publication as a research article in the peer reviewed journal, BMC Genomics and is also on the bioRxiv preprint server:*

Sekar, S., Cuyugan, L., Adkins, J., Geiger, P., & Liang, W. (2018). Circular RNA expression and regulatory network prediction in posterior cingulate astrocytes in elderly subjects. bioRxiv, 268888)

Abstract

Circular RNAs (circRNAs) are a novel class of endogenous, non-coding RNAs that form covalently closed continuous loops and are abundant in the mammalian brain. A role for circRNAs in sponging microRNAs (miRNAs) has been proposed, but the circRNA-miRNA-mRNA interaction networks in human brain cells have not been defined. Therefore, we identified circRNAs in RNA sequencing data previously generated from astrocytes microdissected from the posterior cingulate (PC) of Alzheimer's disease (AD) patients (N=10) and healthy elderly controls (N=10) using four circRNA prediction algorithms. Overall, we identified a union of 4,438 unique circRNAs across all samples, of which 70.3% were derived from exonic regions. Notably, the widely reported CDR1as circRNA was detected in all samples across both groups by find_circ. Given the putative miRNA regulatory function of circRNAs, we identified potential miRNA targets of circRNAs, and further, delineated circRNA-miRNA-mRNA networks using in silico methods. Pathway analysis of the genes regulated by these miRNAs identified significantly enriched immune response pathways, which is consistent with the known function of astrocytes as immune sensors in the brain.

We thus performed circRNA detection on cell-specific transcriptomic data and identified potential circRNA-miRNA-mRNA regulatory networks in PC astrocytes. Given the known function of astrocytes in cerebral innate immunity and our identification of significantly enriched immune response pathways, the circRNAs we identified may be associated with such key functions. While we did not detect recurrent differentially expressed circRNAs in the context of healthy controls or Alzheimer's, we report for the first time circRNAs and their potential regulatory impact in a cell-specific and region-specific manner in aged subjects. These predicted regulatory network and pathway analyses may help provide new insights into transcriptional regulation in the brain.

1. Background

CircRNAs are a class of endogenous, non-coding RNAs that form covalently closed continuous loops and are pervasively expressed in eukaryotes [1-3]. Though RNA circularization events were reported in the 1970s and 1990s [4-6], they were disregarded as molecular artifacts arising from aberrant splicing. However, with the advent of next-generation sequencing technology, coupled with the development of computational algorithms to specifically detect these back-splicing events, numerous circRNAs have been reported since 2012. CircRNAs exhibit cell type-, tissue- and developmental stage-specific expression [3, 7], and show evolutionary conservation between mouse and human [2, 3]. Furthermore, circRNAs are highly abundant in the mammalian brain compared to other tissues such as lungs, heart, kidney, testis and spleen in humans as well as in mouse neuronal cell lines [20], and are derived preferentially from neural genes [22].

The abundance and evolutionary conservation of circRNAs suggests that they could play important roles in cellular processes. A few possible functions have been reported, including microRNA (miRNA) sponges [3, 6, 15, 16], mediation of protein-protein interactions [19] and regulation of parental gene transcription [12]. Furthermore, a few circRNAs have been found to originate from disease-associated genomic loci, suggesting that circRNAs may regulate pathological processes [32, 34, 35, 58, 59]. Given these data, it is likely that circRNAs regulate RNA and protein networks, especially in the brain, but the regulatory pathways are still unknown.

In the present study, we characterized the expression and abundance of circRNAs in next generation RNAseq data of human brain astrocytes. Astrocytes, the most abundant glial cells, play several essential roles in the central nervous system, including homeostasis [60], immunity [61] and energy storage and metabolism [62, 63]. We previously evaluated these astrocytes, which were derived from the posterior cingulate (PC) of Alzheimer's disease (AD) and healthy elderly control brains (age > 65), and identified AD-associated gene expression changes [64]. For this study, we used four circRNA prediction algorithms to identify circRNAs in these AD and control samples. Given the potential miRNA regulatory function of circRNAs, we then performed in silico identification of miRNA binding sites on the detected circRNAs, and further delineated putative circRNA-miRNA-mRNA networks in astrocytes. We describe here the first astrocyte-specific characterization of circRNAs and their interaction networks in elderly individuals.

2. Hypothesis

We hypothesize that circRNAs are differentially expressed in posterior cingulate astrocytes across the AD and ND groups. Given the miRNA regulatory function of circRNAs, we further hypothesize that they regulate miRNAs that in turn regulate mRNAs that are significantly over/under-expressed in astrocytes.

3. Results

3.1 CircRNA detection in PC astrocytes

The RNAseq data generated from our previous study was used for analysis [64]. This data set was generated from 20 human PC astrocyte pools: 10 from late-onset AD (LOAD) brains and 10 from no disease (ND) healthy elderly control brains. Over 85,000,000 reads were sequenced for each sample, with an average mapping percentage of 70.8. On the FASTQ files generated from sequencing, we ran four circRNA prediction algorithms - CIRCexplorer [42], CIRI [46], find_circ [3], and KNIFE [45], and detected a total of 4,438 unique circRNAs with at least two supporting junction reads (publication - Additional file 1: Table S1). Among the detected circRNA candidates, a total of 2,331 circRNAs were identified in the AD samples and 2,425 in the ND samples by at least one of the algorithms (Figure 2.1a). While 80% of the detected circRNAs had less than ten supporting reads (Figure 2.1b), 43 circRNAs had over 20 junction reads and were detected in more than one sample, and 31 circRNA candidates were detected in at least five samples with five or more supporting reads. Notably, the widely reported CDR1as circRNA was detected with a median read count of 52, by find_circ in all 20 samples and by CIRI in one of the samples. CircRNA 2:40655612-40657444 (chromosome:start-end) was detected in 12 of the 20 samples by two, three or all four algorithms in each sample

(publication - Additional file 1: Table S1). Furthermore, 548 circRNAs detected in our dataset were also reported in the four studies deposited in circBase [50] (publication - Additional file 1: Table S1); various cell lines and tissue types were evaluated in these studies, including cerebellum, diencephalon, SH-SY5Y cells, Hs68 cells , HeLa cells and HEK293 cells.

Among all identified circRNAs, 416 were on chromosome 1 (length = 249,250,621 base pairs), while only eight were detected on chromosome Y (length = 59,373,566 base pairs), consistent with previous findings that the number of circRNAs detected is proportional to the length of the chromosome [26]. Based on RefSeq annotations, we observed that 70.3% of our candidates were derived from exonic regions (3,123/4,438), of which 94% (2,936/3,123) were in coding DNA sequences (CDS; excludes untranslated regions) (publication - Additional file 1: Table S1). Among the exonic circRNAs, 56.4% spanned one to 15 exons per circRNA, of which 20% were derived from single exons, while a small percentage of the exonic circRNAs (6.8%) spanned over 100 exons per circRNA.

As previously reported [47], we observed that the overlap among the circRNAs detected by the different tools was low. Overall, 243 circRNAs were predicted by all four tools, while each tool also predicted unique circRNAs (KNIFE—1680, find_circ—1077, CIRI—488, CIRCexplorer—198; Figure 2.1c). Most of the candidates called by all the tools originated from CDS (242/243; 99.5%) as well as intronic regions (232/243; 95.5%), and 75% of the exonic candidates spanned two to six exons per circRNA. Further, the size distribution of all detected circRNAs, and the tool-wise and condition-wise distribution of the circRNAs, are summarized in Figures 2.1d, e and f.

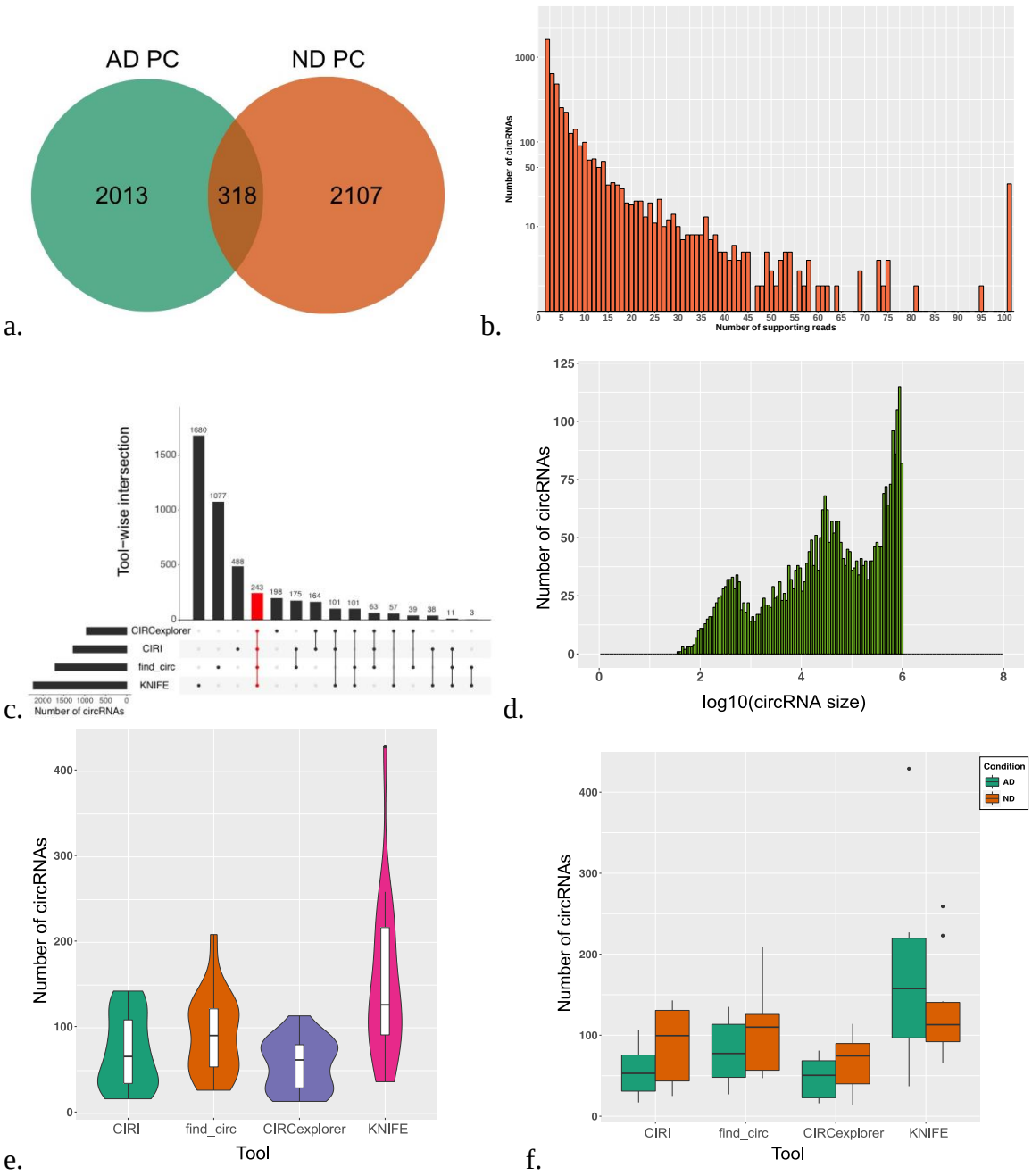


Figure 2.1: Summary of circRNA prediction results. (a) Number of unique and common circRNAs in AD and ND PC. (b) Read count distribution of all detected circRNAs. (c) Intersection of circRNAs called by the four tools; the red bar indicates the number of circRNAs called by all four tools (d) Size distribution of all detected circRNAs. (e) Violin plots indicating the number of circRNAs predicted by each tool across PC samples along with the probability density. (f) Number of circRNAs predicted by each tool across PC samples, condition-wise. AD, Alzheimer's disease; ND, no disease; circRNA, circular RNA; PC, posterior cingulate; bp, base pairs.

We next compared the relative abundance of circRNAs and corresponding linear RNAs using back-spliced reads and linearly spliced reads with the same splice sites (Methods; Table S2; Figure S1). We observed that for 26 circRNAs, the circular-to-linear ratio was 10 or greater and the linear count was not 0, such as circRNA 17:48823196-48824063 from *LUC7L3* (LUC7 like 3 pre-mRNA splicing factor; average back-spliced reads: 413, average linear reads: 16.32) and 1:67356836-67371058 from *WDR78* (WD repeat domain 78; average back-spliced reads: 116.50, average linear reads: 9.50). Further, 44.6% (1,983/4,438) had no expression of linear RNA and 45.5% (2,018/4,438) had higher expression of the linear RNA.

3.2 miRNA target prediction and delineation of circRNA-miRNA-mRNA regulatory networks

Given the potential miRNA regulatory function of circRNAs, we next used the miRNA target prediction algorithms miRanda [65] and RNAHybrid [66] to predict the miRNA targets of the circRNAs detected in ten or more samples by at least one of the circRNA prediction algorithms (N = 10 circRNA candidates). Using a list of 2,588 published miRNAs from miRBase [67], we detected 14,296 unique interactions between circRNAs and miRNAs that were predicted by both the miRNA target prediction algorithms and having a miRanda match score >=150. These interactions represent binding sites for miRNAs on each circRNA candidate, predicted based on complementarity in the miRNA seed region (nucleotide positions 2-7 in the miRNA 5'-end). 2,398 miRNAs in the reference set were predicted to have binding sites on our input list of circRNAs. Among these, a set of 612 circRNA-miRNA interaction pairs were predicted to contain over 100 putative interaction sites by the miRanda algorithm

(publication - Additional file 3: Table S3). These 612 circRNA-miRNA interactions were predicted for six unique circRNAs and 448 unique miRNAs. Using Cytoscape [68], we visualized the circRNA-miRNA interaction network for these 612 interactions, wherein the edges between circRNAs and its target miRNAs are weighted by the number of predicted interaction sites for the circRNA-miRNA pair (Figure 2.2a). CDR1as was predicted to have binding sites for 74 distinct miRNAs and 63 binding sites for miR-7 (Figure 2.2b). According to miRTarBase [69], miR-7 interacts with 578 target genes, some of which include *SNCA* (synuclein alpha), *EIF4E* (eukaryotic translation initiation factor 4E), *KMT5A* (lysine methyltransferase 5A), *MAPKAP1* (mitogen-activated protein kinase associated protein 1), and *MKNK1* (MAP kinase interacting serine/threonine kinase 1).

We further employed the list of miRNA-mRNA target interactions common in both miRTarBase and TargetScan [70] databases, to determine the target genes of the above detected miRNAs. Overall, there were 2,530 target genes for our input list of 2,398 miRNAs, of which 255 were also differentially expressed between the AD and ND groups based on DESeq2 analysis [71] of the linear RNAs (uncorrected $p < 0.05$, publication - Additional file 4: Table S4). Using this information about miRNA target mRNAs, we delineated a putative low-stringency circRNA-miRNA-mRNA network consisting of ten circRNAs, 53 miRNAs and 255 genes (publication - Additional file 9: Figure S2).

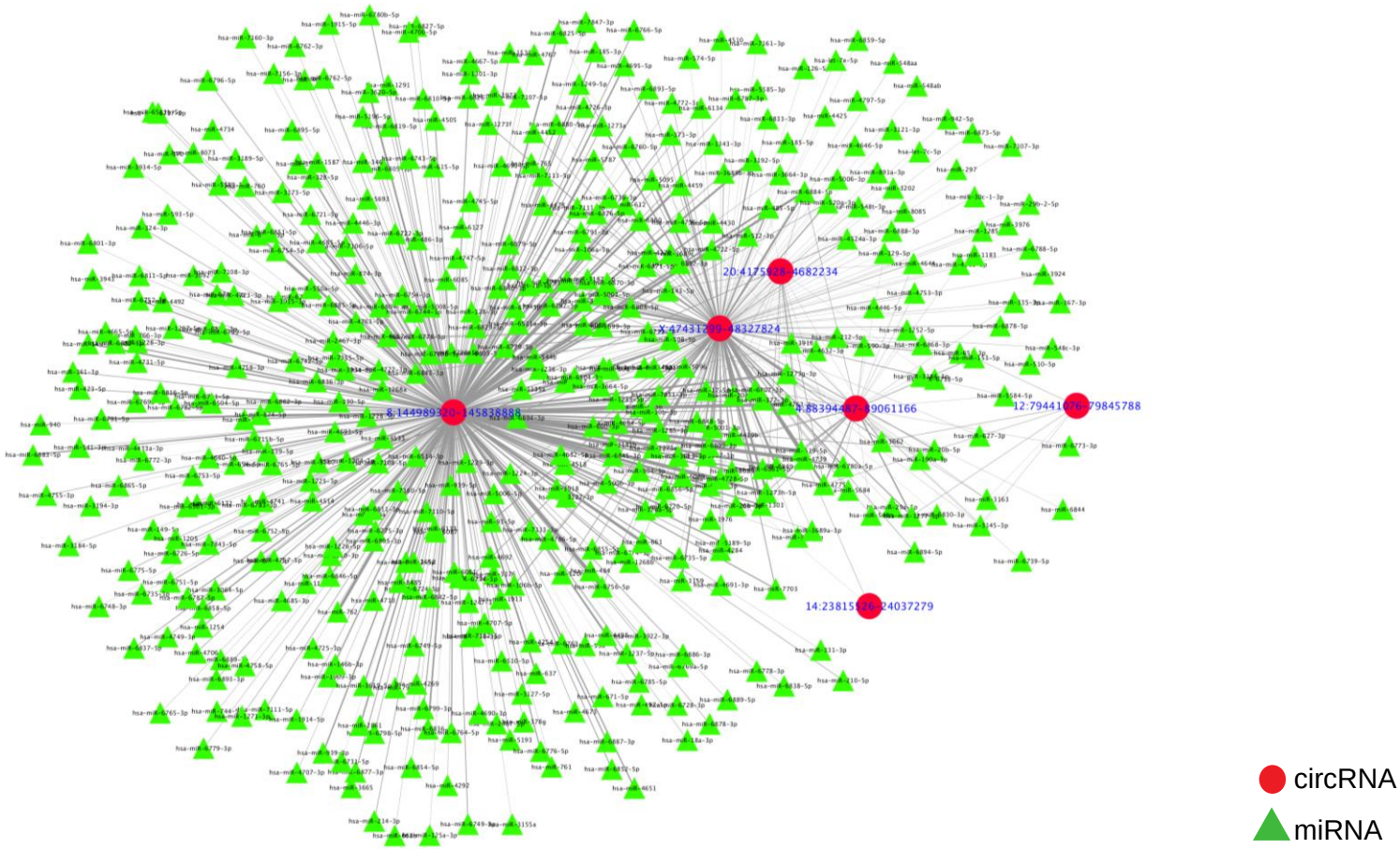
CircRNA	microRNA target	Number of binding sites predicted	Target genes (differentially expressed)
X:47431299-48327824	hsa-miR-139-5p	6	*NOTCH1, STAMBP, TPD52*
8:144989320-145838888	hsa-miR-320a	2	*METTL7A, PBX3, PLS1, SEC14L1, VCL, VIM, VOPP1, YPEL2*
8:144989320-145838888	hsa-miR-320b	2	*RTKN, VCL, VOPP1*
X:47431299-48327824	hsa-miR-449a	1	*BAZ2A, MFSD8, NOTCH1, TSN, ZNF551*
8:144989320-145838888	hsa-miR-125a-3p	1	*ANKRD62, C15orf40, COL18A1, MFSD11, MPEG1, MUL1, TTC31, WDR12, ZNF641*
X:47431299-48327824	hsa-miR-125a-5p	1	*CD34, MEGF9, PANX1, RIT1, TP53INP1*
8:144989320-145838888	hsa-miR-125a-5p	1	*CD34, MEGF9, PANX1, RIT1, TP53INP1*
X:47431299-48327824	hsa-miR-324-5p	1	*FOXO1, MEMO1, PSMD4, SMARCD2*
14:23815526-24037279	hsa-miR-142-3p	1	*BTBD7, CLDN12, CPEB2, CSRP2, DAG1, KIF5B, PTPN23, WHAMM*
4:88394487-89061166	hsa-miR-133b	1	*FAM160B1*
4:88394487-89061166	hsa-miR-448	1	*DDIT4, PURG*
4:88394487-89061166	hsa-miR-339-5p	1	*AXL, HLA-E, METTL7A, ZNF285, ZNRF3*

Table 2.1: circRNA-miRNA-mRNA network elements for those circRNA-miRNA interactions predicted by both miRanda and RNAHybrid, with a miRanda match score >= 180 and mRNA targets differentially expressed (uncorrected $p < 0.05$) with log2(fold change) > 2 or < -2 (high stringency network).

Further, we used the same list of circRNAs detected in ten or more samples by at least one of the circRNA prediction algorithms, and increased the filtering stringency criteria to include a miRanda match score >= 180. We also restricted the candidate miRNAs to those with mRNA targets showing differential gene expression (uncorrected $p < 0.05$) with a log2(fold change) ≥ 2 or ≤ -2 between the AD and ND groups. Using this strategy, we established a high-stringency circRNA-miRNA-mRNA interaction network with four circRNAs, 11 miRNAs and 49 genes (Figure 2.3, Table 2.1). Our overall workflow is outlined in the publication - Additional file 10: Figure S3.

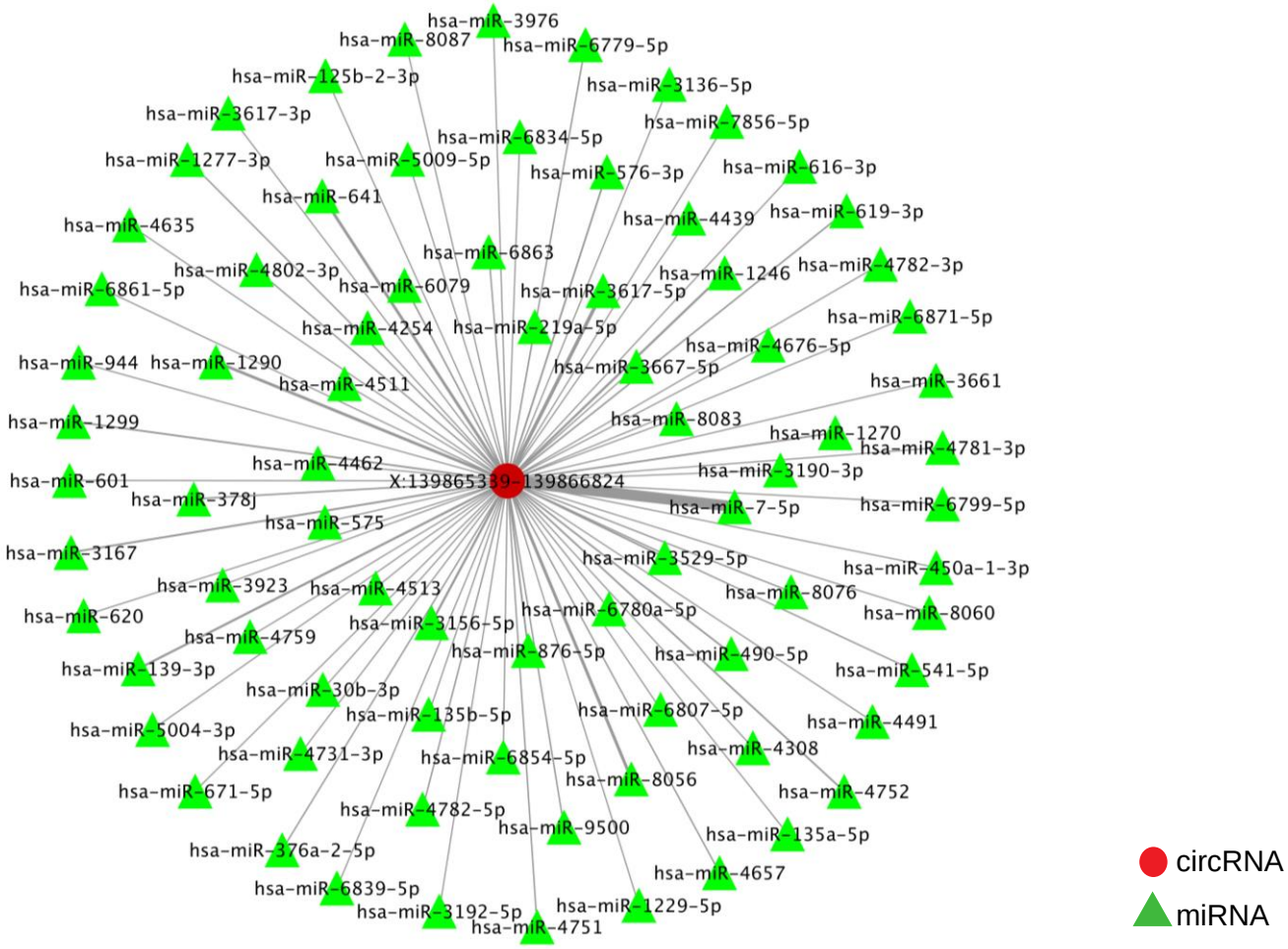
3.3 Pathway analysis

MetaCore pathway analysis on the 255 filtered differentially expressed target genes from the previous analysis revealed 112 perturbed pathways (corrected $P < 0.01$; Appendix Table 2.1, publication - Additional file 5: Table S5). 23 of these were immune response-related, such as IL-4 and IL-6 signaling pathways. This identification of impacted immune response pathways is consistent with the known function of astrocytes as immune sensors in the brain and aligns with our previous RNAseq study, which showed that immune system response pathways are impacted in AD PC astrocytes compared to ND PC astrocytes [64]. Additionally, signal transduction pathways that may be perturbed include post-translational modifications (PTMs) in BAFF-induced signaling, mTORC2 downstream signaling and protein kinase A (PKA) signaling.



a.

Figure 2.2: circRNA-miRNA network. (a) circRNA-miRNA interactions with 100 or more predicted binding sites. Red circular nodes: circRNAs, green triangular nodes: miRNAs.



b.

Figure 2.2: circRNA-miRNA network. (b) miRNA network of CDR1as. The edge thickness in a and b is weighted by the number of binding sites predicted for the circRNA-miRNA interaction. miRNA, micro RNA.

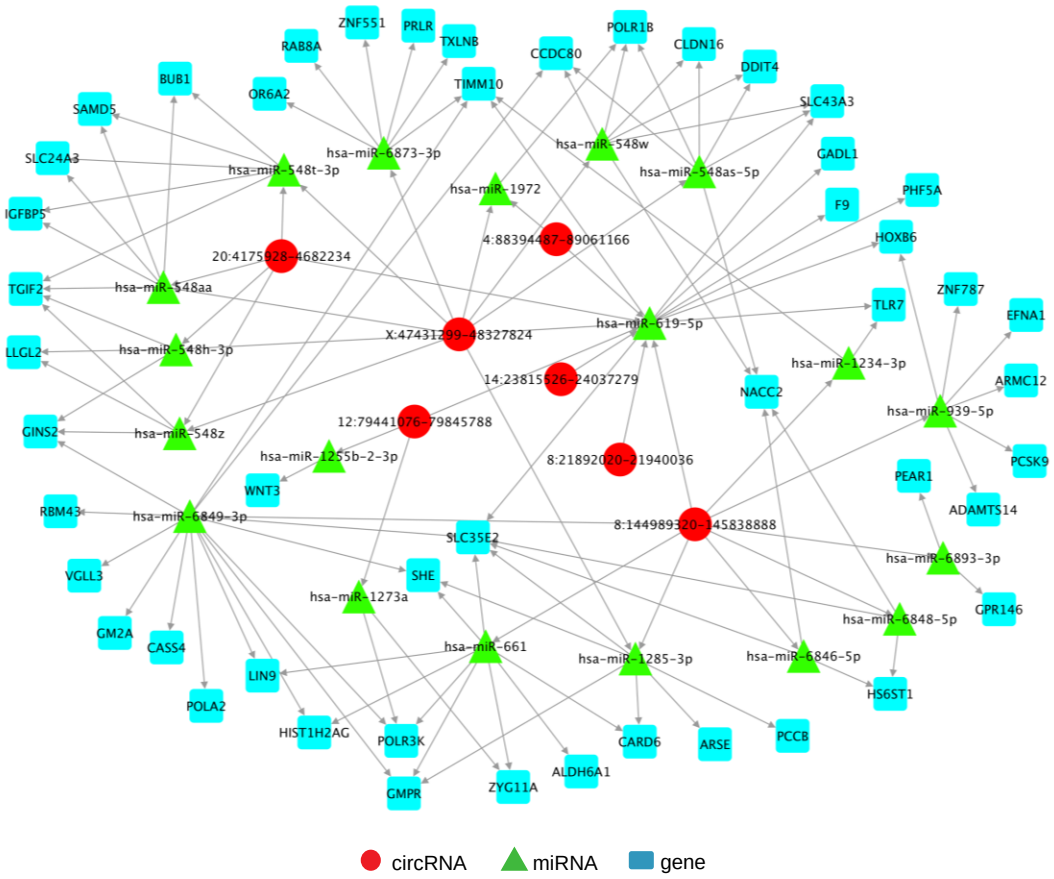


Figure 2.3: High stringency circRNA-miRNA-mRNA regulatory network. Network of circRNA-miRNA-mRNA regulation for those circRNA-miRNA interactions predicted by both RNAHybrid and miRanda, with miRanda match scores >= 180 and mRNA targets with differential expression (uncorrected $p < 0.05$) and log2(fold change) ≥ 2 or ≤ -2. Red circular nodes: circRNAs, green triangular nodes: miRNAs, blue square nodes: genes. mRNA, messenger RNA.

3.4 Lack of circRNA differential expression in AD PC astrocytes

We analyzed our catalog of circRNA candidates to determine whether there were circRNAs uniquely expressed in either the AD or ND cohort. Though there were over 2,000 circRNAs unique to each group, we did not observe them to be recurrent in the samples within their respective group. The log2 (fold change) for all candidates calculated using DESeq2 are summarized in the publication - Additional file 1: Table S1. 93 circRNAs were unique to AD and called in at least two samples by at least one of the tools, and 82 circRNAs were unique to ND and called in at least two samples by at least one of the tools. These circRNA candidates were supported by at least two junction reads. To identify any differentially expressed candidates, we performed a Student's t-test on those circRNAs commonly called across the two groups. Only two circRNAs trending towards significance (uncorrected $p < 0.05$) were identified and include 1:201452657-201736927 (uncorrected $p = 0.015$) and 16:1583657-2204141 (uncorrected $p = 0.046$).

4. Discussion and conclusions

CircRNAs, which are abundant in the mammalian brain, represent a recent addition to the class of non-coding RNAs. In this study, we detected astrocytic circRNAs using whole transcriptome RNAseq data obtained from the PC of AD and ND subjects, and outlined circRNA-miRNA-mRNA regulatory networks. Based on the results from four different circRNA detection algorithms, we identified over 4000 unique circRNAs across all samples, the majority of which were derived from coding exons. Although we did not identify circRNAs that were differentially expressed and recurrent in AD or ND,

we were able to delineate circRNA-miRNA-mRNA networks for the ten most recurrent circRNAs expressed across both groups, and also incorporate our previous differential expression analysis data from the linear mRNA. We observe that the majority of identified circRNAs are unique in the AD or ND groups and are not recurrent across the respective groups. This could be due to their low abundance in those samples, which may be below detection levels, or could be due to biological differences between the two groups, which requires further investigation. Pathway analysis on the differentially expressed miRNA target genes identified immune system related and signal transduction pathways. Notably, astrocytes are active players in cerebral innate immunity [72], and previous studies have reported that astrocytes respond to IL-4 signaling and potentially mediate between the immune effector cells and the nervous responders [73]. These predicted regulatory network and pathway analyses may help provide new insights into transcriptional regulation in the brain.

The circRNA CDR1as (also known as CiRS-7, a circRNA sponge for miR-7) was detected in all 20 of our samples and is a widely reported circRNA with 63 conserved seed matches for miR-7, indicating possible miR-7 binding sites [3, 16]. Interestingly, overexpression of CDR1as in zebrafish decreased the midbrain size, suggesting a functional role for CDR1as in the brain, while a knock down of CDR1as down-regulated miR-7 targets in HEK293 cells [3]. This regulation is relevant since miR-7 plays a role in Parkinson's disease, stress handling and brain development [3, 74], and also has tumor-suppressive properties [74]. CDR1as also showed widespread expression in neuroblastoma and astrocytoma [75]. However, the expression of CDR1as was reduced

in AD hippocampal samples about 0.18-fold compared to controls [32], which we did not observe in our PC astrocyte dataset. Apart from CDR1as, the tools also predicted circRNAs derived from genes such as *SLC8A1* (solute carrier family 8 - sodium/calcium exchanger - member 1), which is under-expressed in hippocampal neurons from aged human brains [76], *SYT1* (synaptotagmin 1), whose increase was correlated to age-related spatial cognitive impairment in mice [77], *PSAP* (prosaposin), which is increased in activated glia during normal aging in mouse brains [78], and *FGF17* (fibroblast growth factor 17).

Although our dataset provides insights into the existence and abundance of astrocytic circRNAs in elderly individuals, there are a few limitations. Primarily, the whole-transcriptome data we analyzed was not generated from samples that were depleted of linear RNAs using RNase R (ribonuclease R), an exoribonuclease that selectively digests linear RNA but leaves behind lariat or circRNA structures. Due to the presence of a larger pool of transcripts, which are mostly linear RNAs, RNAseq may not have comprehensively captured all the circRNAs in the samples. Notably, this enrichment step has been used by various groups to enrich for circRNAs for sequencing analyses [2, 3, 24].

Another limitation of bioinformatics-based circRNA detection is the highly divergent results produced by different algorithms. We observed this in our analyses and it has also been reported by two recent circRNA benchmark studies [47, 48]. The algorithms utilize different aligners, heuristics and filtering criteria, thus introducing ‘blind spots’ (false negatives) when addressing biases introduced by each method [49].

For example, find_circ and CIRI rely on filtering for GT-AG splice signals and thus may not capture candidates with non-canonical splice signals. Further, most tools use a read count filter, which may not be ideal for circRNAs with low expression relative to their linear host [36]. Given the low reliability on read counts, statistical approaches improve detection and classification of splice junctions, including novel ones [79]. Among the circRNA detection algorithms, KNIFE implements a logistic generalized linear model to distinguish true circRNAs, and is therefore able to identify circRNAs derived from non-canonical splice sites. Notably, KNIFE achieves a more balanced performance, for precision and sensitivity, compared to other circRNA detection algorithms, as described in one of the benchmarking studies [48]. We observed in out dataset that KNIFE detected more circRNAs compared to find_circ, CIRI and CIRCexplorer. Nonetheless, sequencing errors and technical artifacts introduced during RNAseq can still lead to false positive circRNAs, and hence statistical tests to estimate false discovery rates in circRNA detection need to be developed.

While circRNAs have continued to gain attention as an abundant non-coding RNA species with potential regulatory functions, our understanding of their expression in various cell and tissue types remains limited. To address this challenge, we describe an analysis of astrocytic circRNAs in RNAseq data from elderly individuals, and we delineate potential circRNA-miRNA-mRNA regulatory networks. Given the role of astrocytes in signaling and synaptic modulation, and as immune sensors in the brain, the circRNAs we identified may be associated with such key functions. Further characterization using circRNA-enriched datasets will help us understand the atlas of

circRNA expression in the context of specific cell types and conditions, including aging and AD. In addition, downstream functional studies are needed to clarify how and whether circRNAs act as hubs for influencing protein expression and cellular processes. As we continue to piece together the factors involved in transcriptional regulation, we will both better understand basic cellular mechanisms and set the stage for developing improved therapeutic strategies for AD and other diseases.

5. Methods

5.1 Sample acquisition, library preparation and paired-end sequencing

Detailed methods for sample acquisition, immunohistochemistry using an aldehyde dehydrogenase 1 family, member L1 (ALDH1L1) antibody, microdissection, RNAseq library preparation and sequencing of astrocytes are described in our previous publication [64]. Briefly, postmortem human brain samples were collected at the Banner Sun Health Research Institute's (BSHRI) Brain and Body Donation Program (BBDP) from 10 clinically classified LOAD subjects (4 males and 6 females; 5 APOEε3/4 subjects and 5 APOEε3/3 subjects) and 10 ND controls (6 males and 4 females; 5 APOEε3/4 subjects and 5 APOEε3/3 subjects). All subjects were enrolled in the BSHRI BBDP in Sun City, Arizona, and written informed consent for all aspects of the program, including tissue sharing, was obtained either from the subjects themselves prior to death or from their legally-appointed representative. The protocol and consent for the BBDP was approved by the Western Institutional Review Board (Puyallap, Washington). Clinical and pathological donor demographics are summarized in the publication -

Additional file 6: Table S6. Approximately 300 astrocytes were laser capture microdissected from PC brain sections and total RNA was isolated from the cell lysates, followed by cDNA creation and library generation. Equimolar pools of libraries were sequenced by synthesis on the Illumina HiSeq2000 for paired 83 base pair reads.

5.2 Data analysis

The data analysis workflow is summarized in the publication - Additional file 10: Figure S3. Raw sequencing data, in the form of basecall files (BCLs), were converted to FASTQ format using Illumina's bcl2fastq conversion software and quality checked using FastQC [80]. To eliminate variance in circRNA detection that could arise due to differences in the number of sequencing reads, all FASTQ files were down-sampled to 85,547,262 reads using seqtk [81]. The down-sampled FASTQ files were then run through four different circRNA prediction algorithms—CIRCexplorer (v1.1.10), CIRI (v2), find_circ (v1), and KNIFE (v1.4), using the parameter settings described in the publication - Additional file 7: Table S7. CircRNAs from each sample with at least two supporting reads were used for further downstream processing and analyses. CIRI produces 1-based circRNA coordinates, and was therefore converted to 0-based coordinates to be consistent with the other three algorithms. We then annotated our catalog of circRNA candidates using the UCSC RefSeq annotations [82] and BEDtools [83].

The ratio of circular to linear RNA isoforms was calculated using the approach described in [20]. For each circRNA candidate, we used the number of back-spliced reads

for circRNA quantification (N_c) and the number of linear reads supporting the same 5' or 3' splice junction (N_{l5} or N_{l3}) as the number of linear RNA reads. The linear junction supporting reads were obtained by aligning our RNAseq data to the reference genome (GRCh37) using STAR [40].

$$\text{Circular to linear ratio } = N_c/\max(N_{l5}, N_{l3})$$

5.2.1 miRNA target prediction

For circRNAs detected in at least 50% of the samples, we next conducted miRNA binding site prediction using the miRanda [65] and RNAHybrid [66] algorithms. The miRanda algorithm finds potential target sites for miRNAs in a genomic sequence using a two-step strategy. First, a dynamic programming local alignment is implemented between the miRNA sequence and the sequence of interest (circRNA sequence in this study), scoring the alignment based on sequence complementarity (match score). In the second step, the thermodynamic stability of the resulting RNA duplex is estimated based on the high-scoring alignments from the first phase. The RNAHybrid algorithm finds the energetically most favorable hybridizations of a small RNA to a large RNA. Only those circRNA-miRNA interactions predicted by both the algorithms are used for our downstream network construction and analyses. From the list of commonly predicted circRNA-miRNA interactions, we filtered for those having a miRanda match score >= 150.

5.2.2 circRNA-miRNA-mRNA network construction

miRNA-mRNA interactions that are common in both miRTarBase [69] and TargetScan [70] were then used to determine the gene targets of each filtered miRNA and

compared with genes identified using differential expression analysis of the linear RNAs (uncorrected $p < 0.05$; DESeq2 performed as described in our previous publication). Using these data, we outlined a low-stringency circRNA-miRNA-mRNA regulatory network with custom python scripts and visualized the network using cytoscape. We further filtered for the circRNA-miRNA interactions with miRanda match scores >= 180 and miRNAs with mRNA targets showing differential expression (uncorrected $p < 0.05$, log2[fold change] ≥ 2 or ≤ -2) to outline a high-stringency circRNA-miRNA-mRNA network.

5.2.3 Pathway analysis

On the list of filtered miRNA target genes with DESeq2 uncorrected $p < 0.05$, we performed pathway analysis using MetaCore GeneGO (v6.32.69020) from Thompson Reuters to predict pathways that are commonly impacted in the AD and ND groups. The results were filtered for enriched pathways with a false discovery rate (FDR)-corrected $P < 0.01$.

6. Ethics approval and consent to participate

All subjects were enrolled in the BSHRI BBDP in Sun City, Arizona, and written informed consent for all aspects of the program, including tissue sharing, was obtained either from the subjects themselves prior to death or from their legally-appointed representative. The protocol and consent for the BBDP was approved by the Western Institutional Review Board (Puyallap, Washington).

7. Availability of data and materials

All the RNAseq data generated in this study are accessible through the National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP; accession# phs000745.v1.p1), and data supporting our findings are included within the manuscript and additional figures/tables.

8. Funding

Research reported in this publication was supported by the National Institute on Aging (NIA) of the National Institutes of Health under award number P30AG019610, and the Arizona Department of Health Services award number ADHS14-052688. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

9. Acknowledgements

We are grateful to the Banner Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona for the provision of human brain tissues. The BBDP has been supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission

(contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research [84]. We would also like to thank TGen's Dr. Kendall Jensen and Dr. Elizabeth Hutchins for input and guidance, and Cynthia Lechuga for administrative support. Nancy Linford, PhD, provided editorial suggestions.

10. Summary

In summary, we utilized astrocyte specific RNAseq data to identify astrocytic circRNAs in aged subjects (N=20). Utilizing four circRNA prediction algorithms, we identified a total of 4,438 unique circRNAs across samples, majority of which were derived from exonic regions. The widely reported CDR1as circRNA was detected in all 20 samples with a median of 52 supporting reads. Although we did not identify circRNAs that were differentially expressed and recurrent in AD or ND, we performed in silico prediction of putative miRNA binding sites on circRNAs detected in at least half the samples, and delineated a low- and high-stringency circRNA-miRNA-mRNA regulatory network. Pathway analysis on the genes from our low-stringency network revealed significantly impacted immune response pathways, which aligns with the known function of astrocytes as immune sensors in the brain. While we did not detect circRNAs recurrently expressed in the context of healthy controls or Alzheimer's, we are the first to report circRNAs and their potential regulatory impact in a cell-specific and region-specific manner in aged subjects. Continued analyses such as these sets the foundation for circRNA characterization and understanding their expression and regulatory networks in specific cell types and regions in the brain.

CHAPTER 3

NOVEL ASSEMBLY BASED APPROACHES FOR FULL LENGTH DETECTION AND *IN SILICO* VALIDATION OF CIRCULAR RNAS

Abstract

CircRNAs are formed when exons 'back-splice' to each other in a non-linear fashion. Although several algorithms have been developed in order to detect these back-splicing events, these methods are limited by a few caveats, which we try to address in this chapter. Firstly, the existing algorithms only look for the presence of the junction and hence we do not detect the full length of the circRNA. We developed DeFuCir (Detection of Full length CircRNA), an assembly based bioinformatics workflow for full length circRNA detection. DeFuCir leverages the presence of soft-clipping signals in the alignment between assembled contigs and the reference genome, and further looks for back-splicing in such contigs to detect potential full length circRNAs. Secondly, existing algorithms produce divergent results, and hence there is a need for an *in silico* validation strategy to distinguish true positive circRNAs. To this end we developed ACValidator (Assembly based CircRNA validator) that can be used as an in silico validation strategy as well as a candidate selection tool for experimental validation. ACValidator extracts reads from a fixed window on either side of the circRNA junction of interest and assembles them to generate contigs. These contigs are then checked for overlap with the circRNAs sequence to check for overlap across the junction. Both these assembly based methods achieve reasonable precision with a higher coverage of back-spliced reads and bring a new perspective to circRNA detection methods.

1. Background

Circular RNAs (circRNAs) represent a large class of ubiquitously expressed non-coding RNAs that are formed when exons 'back-splice' to each other in a non-linear fashion. With the advent of high throughput sequencing technologies and bioinformatics algorithms, thousands of circRNAs have been reported in multiple cell and tissue types by various studies. Notably, these studies have found that circRNAs are highly abundant, evolutionarily conserved and are more stable than linear RNAs since they are covalently closed loops without a 5'/3' termini or a polyadenylated tail. Hence, there has been growing interest in the functional relevance of these RNAs. Nonetheless, recent studies have demonstrated that they can act as miRNA regulators, decoys to RNA binding proteins, and as well as regulators of parental gene transcription.

Several computational tools have been developed to identify these back-splicing events in RNAseq data (Chapter 1, Table 1.2). There are two broad strategies used by these tools to identify circRNAs: 1) a pseudo-reference based strategy used by KNIFE; and 2) a fragment-based strategy used by find_circ, CIRCexplorer, Mapsplice and DCC. While KNIFE constructs a pseudo-reference of all possible out of order exons to align reads against, fragment-based strategies detect circRNAs based on the mapping information of a split read's alignment to the reference genome. When segments of a split read align to the reference in a non-colinear order, they are marked as potential circRNA candidates. Apart from these, CIRI uses CIGAR signatures in the alignment file to identify circRNAs. All of these strategies are alignment based and only look for the presence of the circRNA junction. However, identification of full length circRNAs is

important for understanding their functional relevance. Further, recent comparison studies [47, 48] have revealed that these algorithms produce divergent results, as has also been observed in our results in Chapter 2. Hence, there is a need for an *in silico* validation approach that can help distinguish true versus false positive circRNAs identified using these algorithms.

Previous structural variant detection tools for DNA sequencing data, such as FACTERA [85], CREST [86] and ScanIndel [87], leverage the presence of soft-clipping signals in alignment files to detect indels or translocations. These tools are based on the fact that clipped boundaries of truncated reads (also known as 'soft-clipped' reads) may represent potential DNA breakpoints whereby the soft-clipped portion of a read matches the mapped portion of its mate. Here, we apply a similar principle to utilize soft-clipping signals in order to detect circRNAs and additionally use assembled contigs to predict full length circRNA sequences.

We present two novel computational methods in this study: for the first method, we present a *de novo* assembly based strategy to detect full length circRNA transcripts. This approach utilizes contigs assembled by trinity [88] and identifies soft-clipped bases in the assembled contigs. The advantage of our approach over existing alignment based strategies is the ability to detect essentially full length circRNAs. For the second method, we present a novel *in silico* approach to validate a given circRNA junction of interest. This approach also utilizes the trinity assembler, which assembles reads on either side of the circRNA junction of interest from an alignment file. Generation of a contig that crosses the junction is considered evidence for the presence of a circRNA. This approach

is especially useful in distinguishing true positive candidates as well as in selecting potential candidates for experimental validations. We present a systematic evaluation of our approaches using simulated circRNA datasets generated using CIRIsimulator [46] as well as real datasets consisting of ribonuclease R (RNase R) treated and non-treated sample pairs.

2. Hypothesis

We hypothesize that using an assembly approach can construct potential full length circRNA transcripts. Further, we also hypothesize that assembling reads near the circRNA junction can generate contigs that cross over the junction and can hence be used as an *in silico* validation strategy.

3. Methods

3.1 Assembly based detection of full length circRNAs

We developed DeFuCir (Detection of Full length CircRNAs using assembly and soft-clipping), an assembly based bioinformatics workflow for circRNA detection (Figure 3.1). The first step of our pipeline assembles transcripts using all RNAseq reads from our sample of interest using trinity. Once trinity assembles contigs that are representative of full length transcripts, a series of rules to filter for potential circRNA transcripts among all assembled transcripts is implemented. During this step, the output from trinity, which has the assembled contigs in the form of a fasta file, is first aligned to the reference genome (GRCh37) using BWA-MEM [38]. The resulting binary alignment mapping

(BAM) file is run through a custom python script that inspects each contig's alignment record and collects its start and end coordinates. Further, a second script reads through each alignment record in the BAM file and extracts contigs that have a soft-clipping signal in its CIGAR string. For contigs having a soft-clipping signal, the script extracts the sequence of the lowest and the highest segments of the contig aligning to the reference, lowest and highest being the 5' most and 3' most segments of the contigs alignment to the reference. The tool next checks if the lowest segment of the contig and the soft-clipped portion of the highest segment of the contig overlap with each other. If the lowest segment matches in its entirety to the soft-clipped portion, this is considered as evidence for a back-spliced transcript and thus annotates the contig as a circRNA. Further, the tool also checks if the reverse or the reverse complementary sequence of the lowest segment overlaps with the soft-clipped portion to further check for back-splicing events. Lastly, contigs passing our filtering criteria are collected and their start and stop positions are extracted to generate the predicted circRNA coordinates. When comparing the circRNA coordinates generated from DeFuCir with existing approaches, we allow a 10 base pair (bp) wiggle room to allow for slightly shifted alignment boundaries. Lastly, using a separate parsing script and samtools, the tool also extracts reads aligning to the assembled circular transcript to count how many reads support the circular contig.

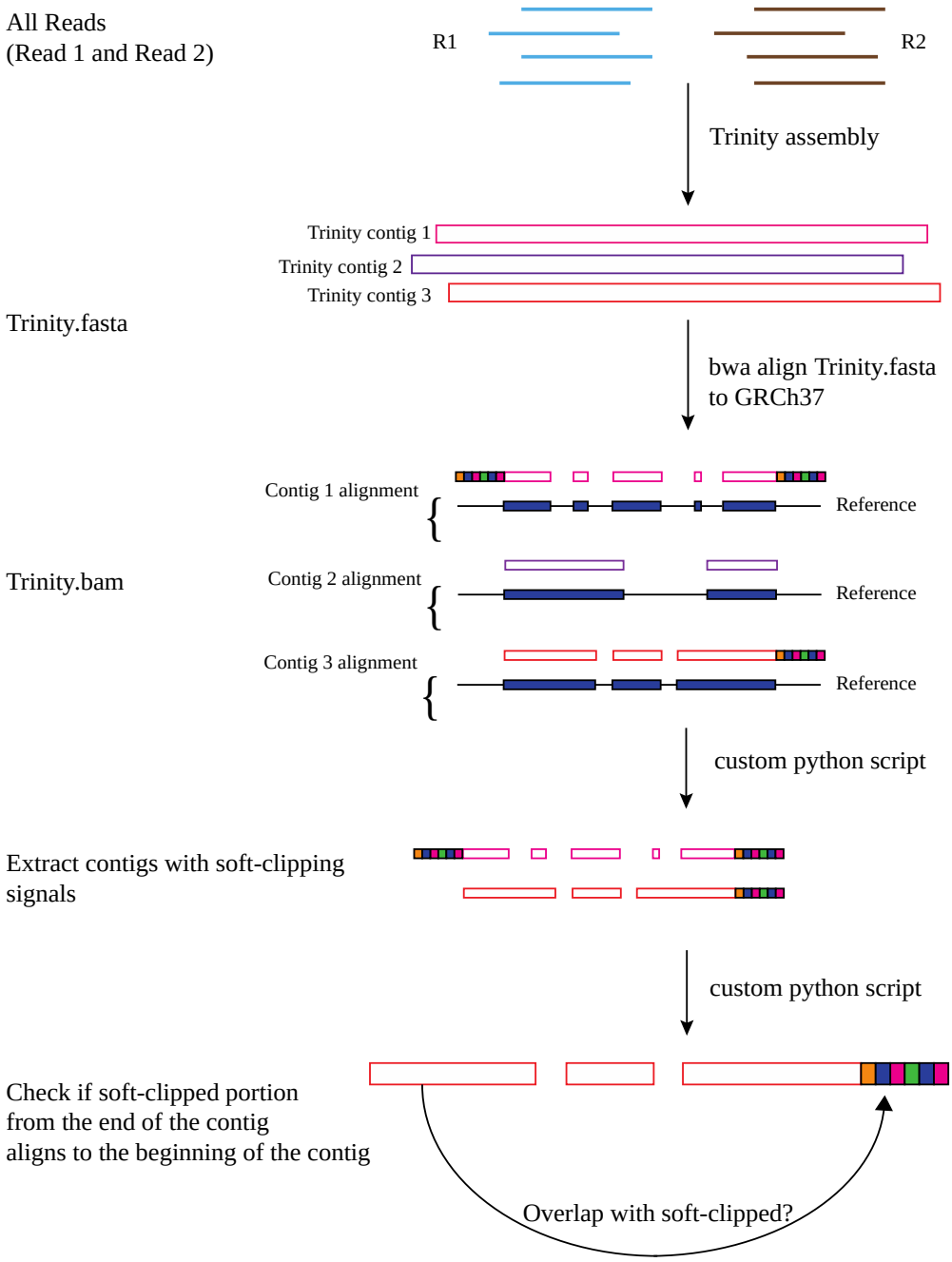


Figure 3.1: DeFuCir workflow. RNAseq reads are assembled using trinity and the resulting fasta file is aligned to the reference genome (GRCh37). Among these alignments, those that have soft-clipping signals at the end of a contig (contig alignments 1 and 3; soft-clipping indicated by the colored bases at the 5' and 3' ends) are further investigated to determine if they map back to the beginning sequence of the same contig. R1, Read 1; R2, Read2

3.2 Assembly based in silico validation of circRNAs

In order to perform an *in silico* validation of selected circRNA junction(s), we developed a bioinformatics workflow, ACValidator (Assembly based CircRNA Validator) that takes as input a SAM file and the circRNA coordinate(s) to be validated (Figure 3.2). ACValidator operates in three phases: extraction of reads from SAM file, generation of a "pseudo-reference" file, and assembly and alignment of extracted reads. First, reads are extracted from a user-defined window *w* on either side of the given SAM file [(*start-coordinate + w)*; (*end-coordinate - w*)]. Our datasets were run using a window size of 300 bp but users can adjust this window according to the library insert size or read length. The tool thus extracts reads aligning between the proximal end of the start coordinate window to the distal end of the end coordinate window from the SAM file using samtools. The extracted reads are converted into fastqs for downstream processing. In the second phase, a "pseudo-reference" of the sequence around the circRNA junction of interest is generated. This is performed by also extracting *w* bps from the end and start of the circRNA junction from the genome reference fasta file (GRCh37) and concatenating the two sequences from end to end to capture the region on either side of the circular junction. Lastly, fastqs from phase 1 are assembled using the trinity assembler and the assembled contigs fasta file is aligned to the pseudo-reference from phase 2 using BWA-MEM. Each resulting alignment record is then examined to check whether they overlap with the junction sequence using four separate stringency criteria. A high stringency cut off requires an overlap of 30 bp on either side of the junction (total 60bp overlap). A 20 bp overlap on either side of the circRNA junction (total 40bp

overlap) is defined as a medium stringency cut off, a 10 bp overlap on either side (total 20 bp overlap) is defined as a low stringency cut off and a 5 bp overlap on either side (total 10 bp overlap) is defined as a very low stringency cut off.

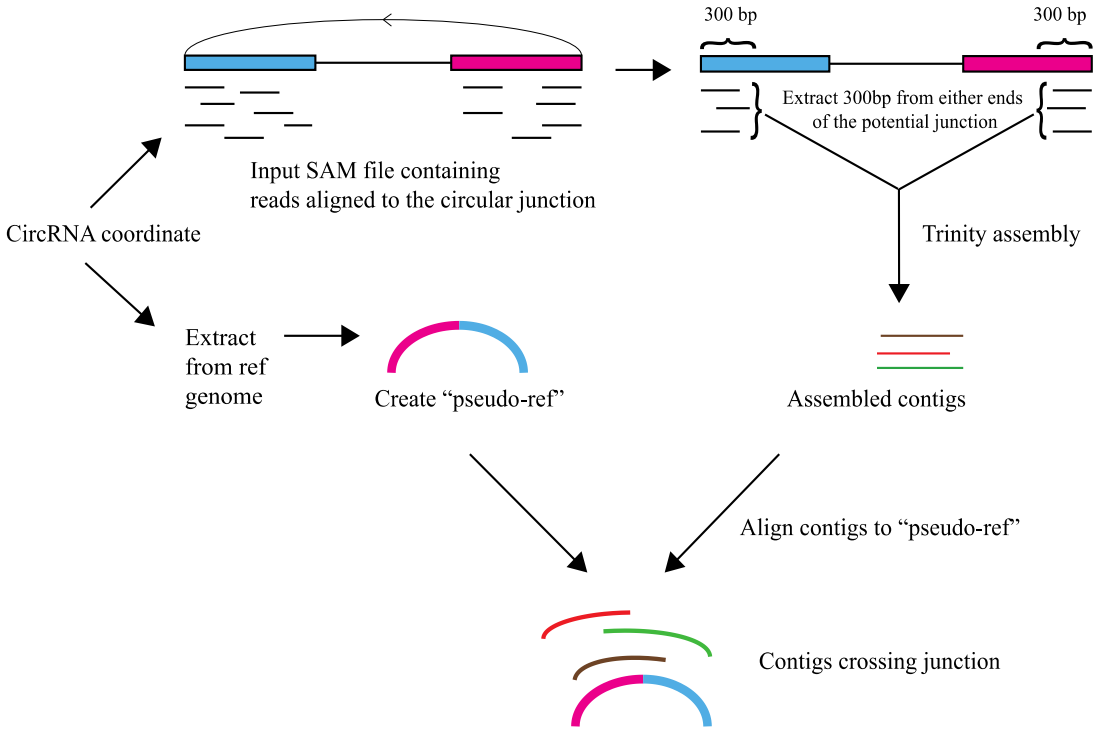


Figure 3.2: ACValidator workflow. ACValidator takes as input the alignment file and the circRNA junction(s) to be validated. Reads from either side of the junction within a user-defined window are extracted and assembled using trinity. The assembled contigs are then checked for overlap with the sequence of the junction to be validated.

3.3 Datasets used for evaluation

3.3.1 Simulated dataset

We used CIRI-simulator [46] to generate five synthetic RNAseq datasets that had varying coverages of circular and linear RNA (Table 3.1) to evaluate our workflows. CIRI-simulator takes as input a fasta formatted reference file and a GTF annotation file based on which it generates circular and linear RNA sequences. Additionally, users can

also specify parameters such as read length, minimum circRNA size, insert size, sequencing error rate, etc. Recently, Zeng *et al*. [48] re-designed this tool to generate synthetic reads for circRNAs deposited in circBase rather than randomly joining exons as per the original design of the tool. The simulated datasets we generated for this study had between two to 24 supporting reads and a minimum circRNA size of at least 50 bps. Overall, there was an average of 56,319 circRNAs across these five simulated datasets. Further CIRI-simulator ensures these circRNAs map to locations distributed across the entire genome, thereby eliminating any bias that may be associated with genomic location (Figure 3.3).

Simulation set	# of reads	CircRNA coverage	Linear RNA coverage	Cell line/tissue type used from circBase	Read length (bp)	Min size of circle (bp)	Insert size (bp)	# True circles
1	2,142,226	10	0	HeLa only	101	50	350	14,689
2	28,920,516	10	0	All*	101	50	300	89,293
3	19,648,436	5	5	All	82	200	300	82,010
4	7,809,356	2	8	All	82	200	300	82,010
5	1,801,264	1	10	HeLa only	82	200	300	13,591

Table 3.1: Simulation dataset parameters

All* cell line/tissue type includes data generated from human cerebellum, diencephalon, SH-SY5Y cells, Hs68 cells, HeLa cells and HEK293 cells

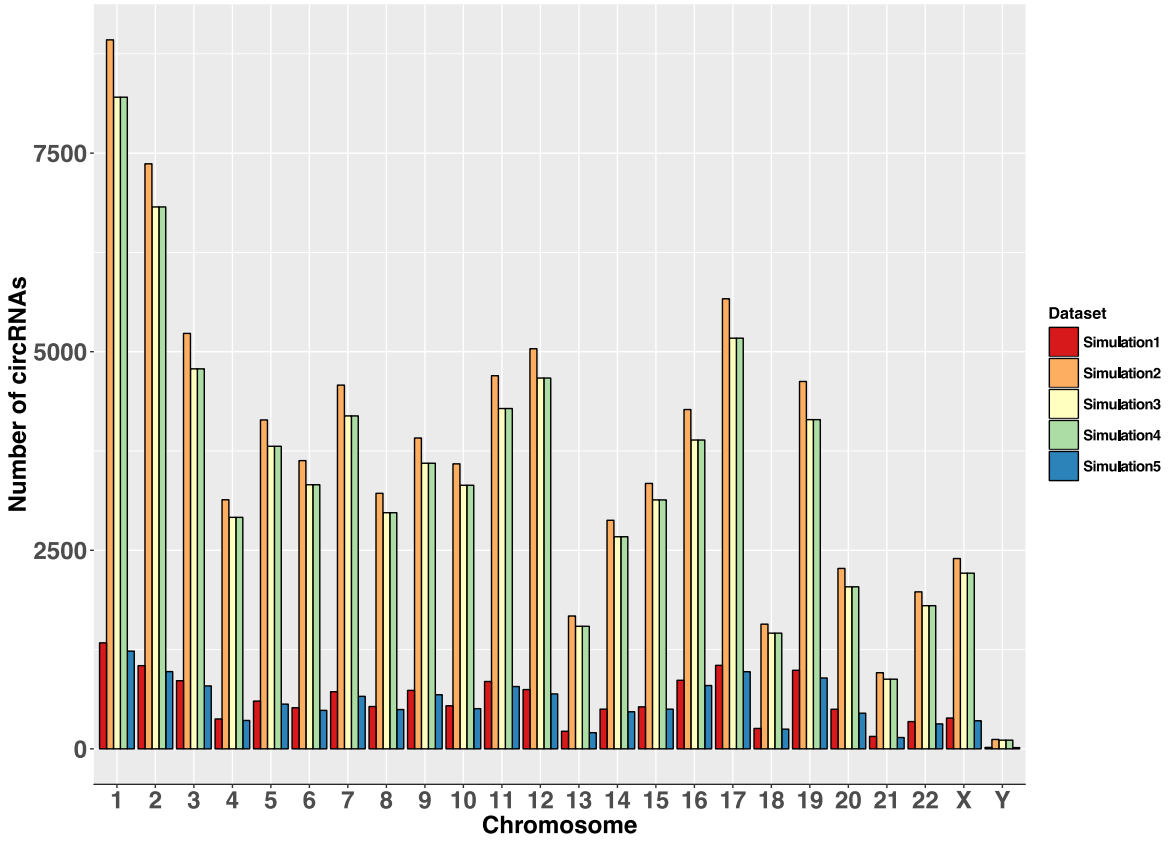


Figure 3.3: Chromosomal distribution of simulated datasets

3.3.2 Non-simulated datasets

We included six pairs of RNase R treated and non-treated samples (N = 12). Three of these were downloaded from the sequence read archive (SRA) [89] and were generated from HeLa and Hs68 cell lines that were treated or not treated with RNase R. The remaining three samples were generated in-house in the Liang lab, TGen, from total RNA extracted from the middle temporal gyrus (MG) of human brain (Chapter 4, Methods). Sample information is summarized in Table 3.2.

Data source	Sample ID	Cell line/ tissue type	RNase R treated ?	Total # reads	# mapped reads
SRA	SRR1636985	HeLa	Yes	26,619,490	24,370,337
	SRR1637089	HeLa	No	89,866,900	63,842,205
	SRR1636986	HeLa	Yes	47,011,426	42,027,801
	SRR1637090	HeLa	No	71,370,620	53,957,685
	SRR444974	Hs68	Yes	316,611,710	271,345,091
	SRR444655	Hs68	No	314,106,316	109,706,923
In-house	MG_1	MG	Yes	107,609,934	96,300,242
	MG_5	MG	No	96,215,516	86,315,844
	MG_2	MG	Yes	96,840,790	86,560,619
	MG_6	MG	No	101,609,754	90,750,108
	MG_3	MG	Yes	111,576,344	100,264,691
	MG_7	MG	No	111,314,114	98,894,498

Table 3.2: Summary of real, non-simulated datasets used in this study

3.4 Software requirements/dependencies

Both workflows were implemented on a Linux-based high performance computing cluster and have minimal requirements and dependencies. These include: (a) Trinity v2.3.1 or above; (b) Python v2.7.13 or higher with pysam package installed; and (c) Bowtie2 v2.3.0, Samtools v1.4, BWA v0.7.12.

4. Results and Discussion

4.1 Assembly based detection of full length circRNAs

We developed DeFuCir, a rule-based bioinformatics workflow for the *de novo* detection of full length circRNAs from trinity assembled contigs. Unlike existing alignment based methods that look for the presence of a back-spliced junction (Chapter 1), this workflow detects full length circRNAs from assembled transcripts based on the presence of soft-clipping signals (Methods). We evaluated our workflow on both simulated and non-simulated datasets.

We first evaluated the performance of our soft-clipping workflow using simulated datasets (Table 3.1). We were able to detect 1,089 to 15,909 candidate circRNAs across the different simulated datasets. Comparing with the true positives in each set as output by CIRI-simulator, our approach achieves reasonable precision because although fewer candidates are detected, the majority of these are true positives (Table 3.3). The main advantage of our workflow over existing methods is that we detect full length circular transcripts, while existing approaches only capture reads at circRNA junctions. This may be especially useful in detecting alternative splicing isoforms whereby certain exons may not be present in the circRNA as inferred (Figure 3.4b).

Sample	# True circles	Total number of contigs from trinity	# contigs with back-spliced soft-clipping signals (circRNAs)	# common circRNAs allowing 10 bp wiggle room	P %
Simulated1	14,689	16,307	2,676	1,827	68
Simulated2	89,293	117,807	15,909	8,706	55
Simulated3	82,010	115,112	10,616	5,638	53
Simulated4	82,010	88,594	6,056	3,285	54
Simulated5	13,591	22,689	1,089	687	63

Table 3.3: Summary of DeFuCir results on simulated datasets. P: Precision, calculated as: TP/TP + FP, where TP: True positives, FP: False positives

We next ran DeFuCir on SRA and in-house generated datasets consisting of RNase R treated and non-treated sample pairs (Methods). When evaluated on the SRA samples, DeFuCir detected 2,476 to 15,076 circRNA contigs. Comparing these results with circRNA candidates identified using existing tools, find_circ, CIRI, Mapsplice, KNIFE, DCC and CIRCexplorer, we detected a total of 4,810 circRNAs across all samples that were common between the two approaches, with an average of 9% overlap per sample (Table 3.4). We observed similar results when implementing our soft-clipping approach for in-house generated MG samples. Overall, DeFuCir detected 11,567 to 24,388 circRNA candidates, among which 191, 154 and 220 candidates overlap between the treated and non-treated pairs. A total of 5,201 circRNAs across all samples overlap between DeFuCir and existing approaches. Among these overlapping candidates, we looked for candidates present in both the treated and non-treated pair to identify a high-confidence list of circRNA candidates that are not depleted by RNase R (Table 3.4). As illustrated in Figure 3.4, candidates detected by DeFuCir span the full length of the circRNA (top panel), whereas existing tools only look for the presence of reads near the

predicted circRNA junction (bottom panel). Furthermore, Zeng et al. [48] curated a list of 282 PCR (polymerase chain reaction) validated circRNAs from various published studies. Though these validated circRNAs are derived from different cell lines and tissue types, we compared our catalog of back-spliced contigs with the validated circRNAs and detected overlapping events (Table 3.4).

Sample	RNase R treated?	# contigs with back-spliced soft-clipping signals (circRNAs)	# overlap across pair	# overlap with existing approaches	Intersection between treated-non-treated pairs in columns 5	# overlap with validated circRNAs (N=282)
MG_1	Yes	17,599	191	1,987	106	20
MG_5	No	11,567		391		13
MG_2	Yes	19,504	154	1,319	77	17
MG_6	No	17,205		348		18
MG_3	Yes	44,518	220	850	78	18
MG_7	No	24,388		306		15
SRR1636985	Yes	2,476	40	423	35	17
SRR1637089	No	12,622		612		8
SRR1636986	Yes	4,086	57	254	43	24
SRR1637090	No	7,467		281		19
SRR444974	Yes	15,076	90	3,148	35	26
SRR444655	No	7,494		92		7

Table 3.4: Summary of results from running DeFuCir on non-simulated datasets

'# overlap' indicates 'number of candidates overlapping'; validated circRNAs from Zeng et al 2017

a.

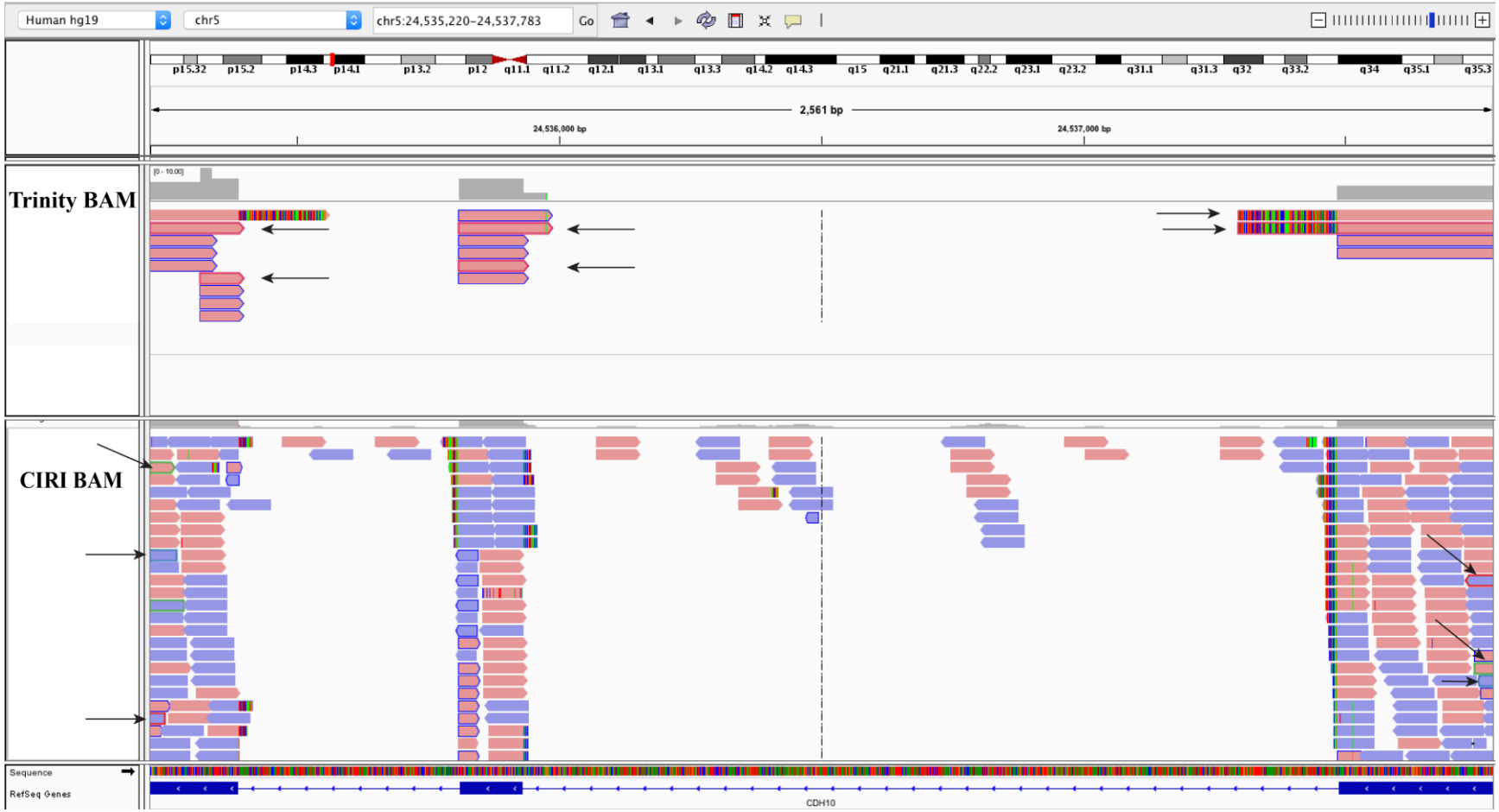


Figure 3.4: Integrated genomics viewer (IGV) screen shots of two circRNA candidates detected by DeFuCir and existing tools (panel a shown)

b.

Figure 3.4: Integrated genomics viewer (IGV) screen shots of two circRNA candidates detected by DeFuCir and existing tools (a. 5:24,535,220-24,537,783; b. 10:105,953,623-105,967,582). The top panel in each screenshot is the BAM file from DeFuCir while the bottom panel is from running through CIRI and CIRCexplorer in these examples. Arrow marks indicate contigs/reads supporting the circRNA transcript/junction. We observe that DeFuCir generates contigs spanning the exons constituting the circRNA transcript and in (b) there are no contigs over exons 2 and 3 indicating that there may be alternative splicing, which cannot be inferred from the CIRCexplorer BAM file.

4.2 Assembly based in silico validation of circRNAs

Given the disparity among circRNAs identified by existing circRNA algorithms, we implemented an assembly based workflow for *in silico* validation of circRNAs using ACValidator (Methods). Briefly, ACValidator takes as input the circRNA junction(s) to be validated and the alignment file. Using a user-defined window, reads from either side of the junction are extracted and assembled using trinity. The assembled contigs are then checked for overlap with the sequence of the junction to be validated.

To evaluate ACValidator, we first used the same five simulation datasets described above. We used the top 100 true candidates in terms of the highest number of read counts as true positives, and randomly selected non-circRNA coordinates as false positive candidates. As summarized in Table 3.5, our workflow achieves higher sensitivity and precision on datasets with a higher number of circRNA supporting reads and gradually reduces with circRNA coverage. In most cases, this is due to the fact that trinity does not identify sufficient reads to assemble across these regions and hence does not generate contigs. In summary, simulations 1 and 2 achieve 92 and 83% sensitivity respectively while the remaining 3 datasets achieve relatively lower sensitivity. Further, we calculated the F1 score [F1 = (2 * Precision * Sensitivity)/ (Precision + Sensitivity)], which indicates how well a tool achieves sensitivity and precision simultaneously, and observed the results to be consistent with the sensitivity measurements, indicating that higher circRNA coverage yields better performance of our approach.

Simulation set	# true candidates	# false candidates	TP	FN	FP	P	S	Sp	F1
1	100	100	92	8	1	99	92	99	0.95
2	100	100	83	17	2	98	83	98	0.89
3	100	100	76	24	11	87	76	90	0.81
4	100	100	71	29	1	99	71	99	0.82
5	100	100	60	40	1	98	60	99	0.74

Table 3.5: Results of ACValidator on the top 100 candidates and false candidates
The top 100 candidates are those with the highest number of supporting reads and false candidates are randomly selected non-circRNA coordinates. In order to determine the FP rate, 100 false candidates were also tested to evaluate how many were validated by our tool.
TP, True positives; FN, False negatives; FP, False positives; P, Precision; S, Sensitivity; Sp, Specificity.
P =TP/(TP+FP); S = TP/(TP + FN); Sp = TN/(TN+FP); F1 = (2 * P * S)/ (P + S)

We extended these calculations to the top 200 as well as bottom 200 candidates, as defined by the number of supporting reads and observed the same trend in sensitivity and precision measurements (Table 3.6). Furthermore, we also evaluated sensitivity using different thresholds for the number of overlapping bases (30 bp, 20 bp, 10 bp and 5 bp; Methods) to assess whether we achieve "saturation" in the sensitivity measurements. For majority of the cases, we observe that there is no drastic difference in the number of candidates validated across the different thresholds.

a. Top 200 candidates

Simulation set	HS	MS	LS	VLS	S % (HS)	S % (MS)	S % (LS)	S % (VLS)
1	180	183	183	183	90.00	91.50	91.50	91.50
2	158	163	164	164	79.00	81.50	82.00	82.00
3	149	156	156	156	74.50	78.00	78.00	78.00
4	133	144	146	146	66.50	72.00	73.00	73.00
5	114	121	122	122	57.00	60.50	61.00	61.00

b. Bottom 200 candidates

Simulation set	HS	MS	LS	VLS	S % (HS)	S % (MS)	S % (LS)	S % (VLS)
1	97	109	109	111	48.50	54.50	54.50	55.50
2	69	84	87	93	34.50	42.00	43.50	46.50
3	34	41	42	44	17.00	20.50	21.00	22.00
4	26	30	36	40	13.00	15.00	18.00	20.00
5	40	49	52	53	20.00	24.50	26.00	26.50

Table 3.6: ACValidator results summary for a. top 200 and b. bottom 200 circRNAs in the simulation dataset. The top and bottom 200 candidates are defined in terms of the number of supporting reads. HS, high stringency; MS, medium stringency; LS, low stringency; VLS, very low stringency; S sensitivity

We next evaluated ACValidator using non-simulated datasets generated from tissues or cells and that are summarized in Table 3.2. Since we do not know the true positive candidates for these datasets, we validated those circRNAs that were called in both the RNase R treated and non-treated datasets using six existing circRNA detection algorithms, find_circ, CIRI, Mapsplice, KNIFE, DCC and CIRCexplorer. Using SRPBM values, we selected those candidates that were called by at least three of the six tools, common across the treated and non-treated samples and not depleted of linear RNAs

using RNase R (Chapter 2, Methods). Overall, except for the SRR1636986-SRR1637090 pair, over 89% of the candidates that were common across the treated-non-treated pairs were not depleted. Among these non-depleted candidates, ACValidator was able to construct contigs for more than 75% of them for the RNase R treated samples and 47-56% of the candidates for the non-treated samples (Table 3.7). This increased validation rate for the treated samples is expected since RNase R treatment enriches the majority of the circRNA species and hence they have a higher number of back-splice junction supporting reads. Figure 3.5 shows two examples of circRNAs validated by ACValidator in the treated and non-treated pairs.

Dataset	# overlap with pair	# not depleted	# validated HS	# validated MS	# validated LS	# validated VLS	HS%	MS%	LS%	VLS%
SRR1636985	1356	1243	918	982	990	996	73.85	79.00	79.65	80.13
SRR1637089	1356	1243	559	620	628	642	44.97	49.88	50.52	51.65
SRR1636986	780	594	481	510	516	520	80.98	85.86	86.87	87.54
SRR1637090	780	594	287	328	334	336	48.32	55.22	56.23	56.57
SRR444974	953	864	777	795	799	802	89.93	92.01	92.48	92.82
SRR444655	953	864	461	492	494	509	53.36	56.94	57.18	58.91
MG_1	1806	1691	1212	1308	1319	1331	71.67	77.35	78.00	78.71
MG_5	1806	1691	699	825	840	852	41.34	48.79	49.67	50.38
MG_2	1430	1331	928	1013	1020	1029	69.72	76.11	76.63	77.31
MG_6	1430	1331	537	626	634	649	40.35	47.03	47.63	48.76
MG_3	1292	1148	808	871	882	892	70.38	75.87	76.83	77.70
MG_7	1292	1148	511	590	599	615	44.51	51.39	52.18	53.57

Table 3.7: Summary of ACValidator results on non-simulated datasets. Overlap with pair indicates the number of circRNAs common between the treated and untreated pairs and # not depleted is the count of circRNAs from the overlap whose SRPBM does not decrease following enrichment. HS%, MS%, LS% and VLS% indicate the percentage of candidates validated using high stringency, medium stringency, low stringency and very low stringency cut-offs.

a)

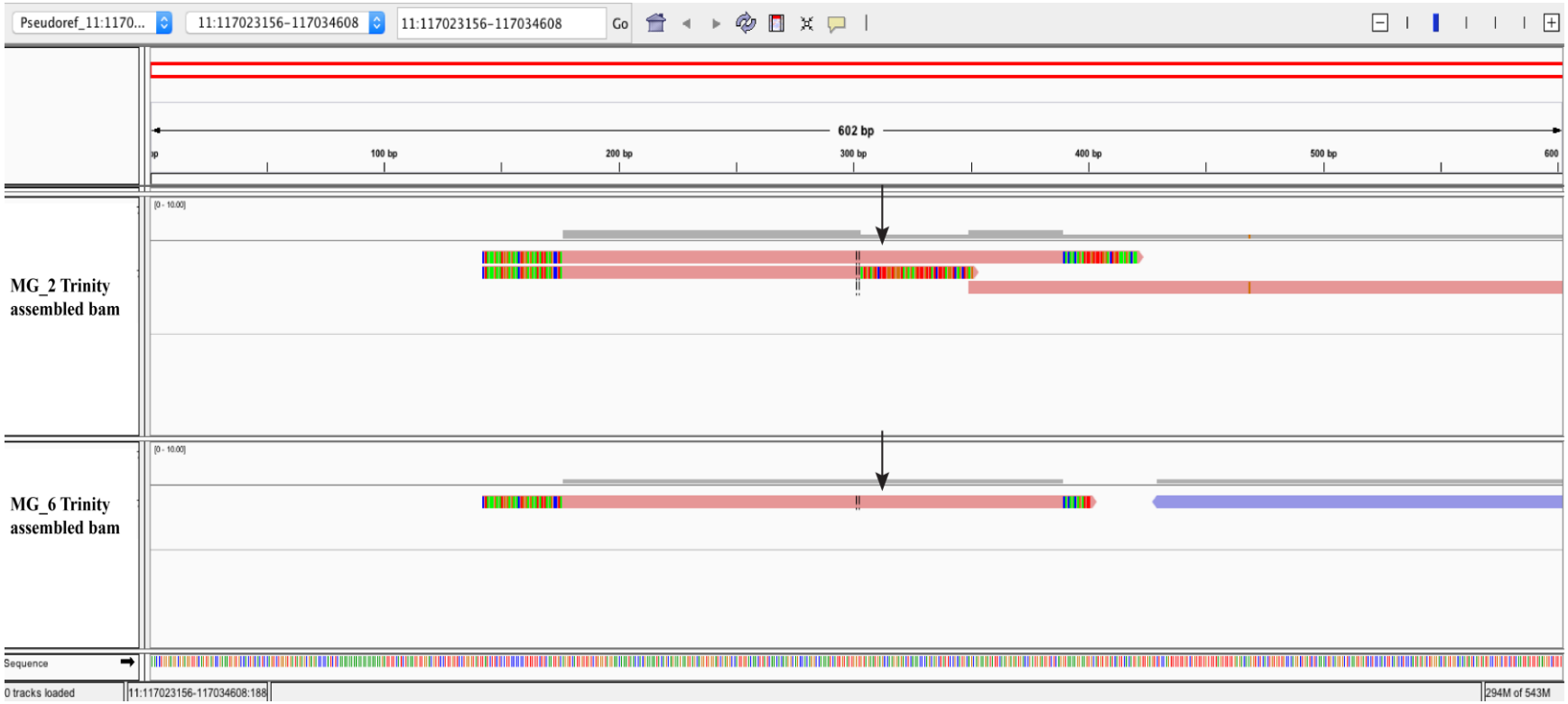


Figure 3.5: IGV screen shots of two circRNA candidates validated by ACValidator on RNAse R treated and non-treated samples (panel a shown)

b)

Figure 3.5: IGV screen shots of two circRNA candidates validated by ACValidator on RNAse R treated and non-treated samples (a. 11:117,023,156-117,034,608 and b. 9:79,829,244-79,842,401) In both screenshots, the top panel is the trinity BAM from the RNAse R treated sample and the bottom panel is the non-treated sample. Arrows indicate the trinity constructed contigs that cross the circRNA junction of interest (the junction itself is at position 300 and are indicated by double lines in the center of the contig, the arrows indicate the contig generated that crosses the junction)

4.3 Computational cost overview

As expected, the computational cost of our approaches directly correlates with the sequencing depth of the input sample. With respect to DeFuCir, the main rate-limiting step is running trinity assembly. We observed that on an eight core high performance computing cluster, trinity ran for an average of two days on files of size 16 GB (read 1+read 2 FASTQs). Following trinity, the custom python post-processing scripts consume less than 10 minutes to complete per sample. As for the validation workflow, the only rate- limiting step is BWA-MEM alignment, which is performed prior to starting this workflow. The python validation script following this step takes less than 2 minutes to run for one sorted BAM file of approximately 8 GB, thus making this approach highly computationally efficient.

4.4 Conclusions

In this study, we present two novel computational workflows DeFuCir and ACValidator, both of which address two informatics challenges associated with existing detection methods. Firstly, though several circRNA detection tools have been developed, these alignment based methods only look for the presence of reads supporting the back-spliced junction. However, obtaining the full length circRNA sequences will be valuable for enabling further functional analyses aimed at elucidating alternative splicing mechanisms in circRNAs. We developed DeFuCir, an assembly based workflow for the detection of full length circRNAs, which leverages soft-clipping signals in an alignment file. Though DeFuCir detects a smaller number of candidates, it achieves improved

precision in that the majority of the candidates it detects are true positives. We thus propose DeFuCir as an approach that can complement existing methods by filtering for high confidence candidates that overlap between the two and by obtaining full length circRNA sequences. In order to implement an unbiased approach to circRNA detection, DeFuCir uses only minimal filtering thresholds, and hence, repeat regions or single nucleotide variants (SNVs) in overlapping regions are currently not accounted for in our pipeline.

Secondly, in order to address the issue of diverse results from these existing tools, we present ACValidator, a novel bioinformatics workflow that can validate candidate circRNAs of interest and help distinguish true positive candidates. When different circRNA detection tools identify different circRNA candidates, ACValidator will be helpful in narrowing down specific candidates of interest. Further, this can also serve as a circRNA candidate selection tool for experimental validations or functional studies. ACValidator achieves better performance with higher circRNA coverage when assessed on simulated datasets. Additionally, when tested on RNAse R treated and non-treated datasets, we observe that the RNAse R treated samples achieved a higher validation rate due to enrichment. The different overlap stringency thresholds help to assess whether validation approaches reach saturation and also helps ensure we capture as many validations as possible, while still accounting for the extent of overlap for each of them.

As discussed in Chapter 2, continued development of novel approaches, including implementation of statistical tests to estimate false discovery rates in circRNA detection, are needed. Further, progress in understanding the biology of circRNAs will be necessary

for such algorithmic development. For example, existing tools do not consider gene fusion circRNAs but recently Guarnerio et al. [90] demonstrated that gene fusion derived circRNAs may play a potential role in cancer pathogenesis. Such findings will be crucial not only for functional analysis, but also in the development of more accurate circRNA detection algorithms that account for such events.

5. Summary

We developed two novel assembly based workflows, DeFuCir and ACValidator that aim to address some of the informatics challenges associated with current circRNA detection methods. DeFuCir is an assembly based workflow that constructs full length circRNA contigs that represent potential full length circRNA sequences based on the presence of back-splicing evidence. Although DeFuCir achieves lower precision compared to existing tools, we propose this as an approach that can complement existing approaches to detect full length ensemble circRNA candidates. ACValidator is an *in silico* validation strategy that can be used to confirm the existence of a circRNA junction based on assembled contigs that cross over the junction. ACValidator achieves high precision and sensitivity with higher circRNA coverage as well as in RNAse R enriched datasets. Further, ACValidator can be used as a candidate selection strategy for experimental validation of circRNAs. These two assembly based approaches introduce a novel perspective to current circRNA detection methodologies.

CHAPTER 4

CIRCULAR RNAS IN FUNCTIONALLY DISTINCT REGIONS OF HEALTHY AGED HUMAN BRAIN

Abstract

CircRNAs, a recently characterized species of non-coding RNAs, are both highly conserved and abundant in the mammalian brain. In this study, we systematically identify and characterize them in circRNA-enriched RNAseq data from five functionally distinct regions of the human brain, including cerebellum (BC), inferior parietal lobe (IP), middle temporal gyrus (MG), occipital cortex (OC) and superior frontal gyrus (SF) (N = 20 total samples, four from each regions). Overall, over 25,000 circRNAs were detected across all the regions, among which 4,528 were identified by at least three of the six circRNA detection tools used, and in all four samples from a region. We also identified differentially expressed circRNAs in each brain region compared to all other regions, for which potential circRNA-miRNA-mRNA regulatory networks were delineated. Ingenuity pathway analysis on genes from these networks identified several nervous system related functions and pathways specific to each region, indicating that circRNAs could be involved in such key functions. Further, we identify a list of high-confidence full length circRNAs which are detected by both DeFuCir as well as exiting tools. Among these, circRNAs from the *RIMS* gene was detected in four of the five brain regions with SRPBM values > 2000. Lastly, using ACValidator, we perform in silico validation of ~45% of differentially expressed circRNAs in each region, and over 80% of circRNAs called by all the existing tools in all samples from a region.

1. Background

Circular RNAs (circRNAs) represent a species of non-coding RNAs that have yet to be well characterized and are formed when a downstream splice donor splices to an upstream splice acceptor. It has been recently demonstrated that they regulate microRNAs (miRNAs), preventing them from binding to their target mRNAs, as well as act as decoys to RBPs. Further, recent studies have reported preferential back splicing of neural genes and an abundance of circRNAs in the mammalian brain. Rybak-Wolf et al. [20] analyzed human and mouse neuronal cell line data and found that circRNAs are highly abundant in the mammalian brain compared to other analyzed tissues such as lungs, heart, kidney, testis and spleen, with well-conserved sequences. They also found that circRNAs were upregulated during neuronal differentiation and development, and highly enriched in synapses, independent of their corresponding linear isoform. In a separate study, analysis of RNAseq data from mouse brain, liver, heart, lung and testis revealed that although circRNAs were present in all the examined tissues, their abundance was highest in the brain. Further, 20% of the protein coding genes in the brain were found to produce circRNAs and encode for proteins involved in several synapse-related functions. The study also found an enrichment of circRNAs compared to their host linear transcripts in synaptosomes and microdissected neuropils from mouse hippocampal slices [22].

Given the high abundance of circRNAs in the brain, there is a growing interest in the role of circRNAs in neurological diseases. Recently, Lukiw [32] reported reduced levels of circRNA ciRS-7 and a dysregulated ciRS-7-miR7 system in the hippocampal

CA1 region of Alzheimer's disease (AD). However, functional studies still need to be performed to elucidate the role of ciRS-7 in AD and other neurological conditions.

In this study, we aim to systematically identify and characterize circRNAs in five functionally distinct regions of healthy aged human brain including the cerebellum (BC), inferior parietal lobe (IP), middle temporal gyrus (MG), occipital cortex (OC) and superior frontal gyrus (SF). We further identify circRNAs from other tissue types including lung (LU), liver (LV), lymph node (LN) and pancreas (PA) and compare their abundance to that of brain circRNAs. Given the miRNA regulatory function of circRNAs, we further aim to evaluate their predicted impact on regulatory networks in each brain region using *in silico* methods.

We thus performed next generation RNA sequencing (RNAseq) of RNase R (ribonuclease R) treated total RNA and describe for the first time brain region specific circRNA expression using circRNA-enriched datasets. Using these analyses, we establish a reference of circRNA expression profiles and regulatory networks in healthy elderly individuals in a region-specific manner. This resource along with existing databases such as circBase will be invaluable in advancing circRNA research and furthering our understanding of their role in transcriptional regulation in the brain and as well as in other neurological processes and conditions.

2. Hypothesis

We hypothesize that there are unique circRNA expression profiles in the five functionally distinct brain regions: BC, IP, MG, OC and SF. Further, we also hypothesize that these

circRNAs regulate miRNAs that in turn regulate mRNAs involved in functions specific to each brain region.

3. Results and discussion

3.1 Overall circRNA detection results

20 RNAseq libraries were sequenced from five brain regions across four separate subjects to generate a median of 98,416,405 reads across all samples. Sequencing summary metrics are summarized in Table 4.1. On the FASTQ files generated from sequencing, we ran six circRNA prediction algorithms and detected a union of 28,220 circRNAs across all samples with at least two supporting reads each (tool-wise breakdown of all detected circRNAs are summarized in Figure 4.1a). Among these, 752, 3496, 355, 1357 and 3478 were called by at least three of the six tools in all four samples in BC, IP, MG, OC and SF, respectively (Figure 4.1b; henceforth referred to as our list of "high confidence" circRNAs). CDR1as, the widely reported circRNA, was detected by two of the tools - DCC and find_circ, with a median of 539 supporting reads across all samples.

The majority of the identified high-confidence circRNAs map to the CDS (coding DNA sequence) and intronic regions, while a small percentage also map to intergenic and 3' untranslated (UTR) regions (Table 4.2). Further, over 88% candidates in all regions contained two to ten exons per circRNA whereas 3% or less had 10 or more exons. CircRNAs 10:116,879,948-117,309,046 (chromosome:start_coordinate-end_coordinate) from gene *ATRNL1* (attractin-like 1) and 2:55,040,368-55,155,888 from gene *EML6*

(echinoderm microtubule associated protein like 6) had the highest number of exons - 26. However, it needs to be noted that these tools focus on junction-supporting reads only and hence may not account for alternative splicing events, as discussed in Chapter 3.

Sample ID	Total reads	Total reads after UMI processing	Mapped reads	Mapped reads %	rRNA %	mRNA %	Intronic reads %	Intergenic reads %
BC_1	110,166,426	107,208,904	96,842,861	90.33	2.27	44.89	43.08	9.82
BC_2	114,951,534	111,833,362	100,749,352	90.09	0.68	33.51	45.74	20.09
BC_3	98,639,608	96,125,856	83,742,083	87.12	0.25	7.98	47.46	44.32
BC_4	100,710,470	96,277,094	80,376,250	83.48	0.15	5.41	49.30	45.15
IP_1	101,386,268	98,656,506	88,348,693	89.55	6.79	48.29	31.74	13.37
IP_2	117,002,834	113,595,746	101,666,810	89.50	3.40	49.60	34.47	12.61
IP_3	132,785,312	128,781,536	113,670,893	88.27	1.88	48.04	28.50	21.63
IP_4	98,769,292	95,981,446	85,452,117	89.03	1.62	33.89	36.57	27.95
MG_1	111,473,936	107,609,934	96,300,242	89.49	3.79	53.06	33.12	10.12
MG_2	99,559,694	96,840,790	86,560,619	89.38	6.17	46.80	33.44	13.75
MG_3	114,053,720	111,576,344	100,264,691	89.86	1.05	24.15	41.35	33.47
MG_4	103,003,950	100,539,028	88,511,086	88.04	0.26	8.11	48.07	43.57
OC_1	90,848,494	88,547,042	80,250,817	90.63	5.22	45.81	37.56	11.54
OC_2	124,200,016	120,983,142	107,386,487	88.76	6.35	49.13	31.99	12.67
OC_3	105,523,894	102,655,452	92,080,615	89.70	4.36	43.00	29.95	22.75
OC_4	112,150,858	108,449,330	94,107,535	86.78	0.22	10.34	47.67	41.78
SF_1	90,538,780	87,621,824	78,534,192	89.63	3.10	52.60	32.88	11.49
SF_2	149,529,412	145,057,894	128,027,756	88.26	6.94	53.86	27.30	12.05
SF_3	112,711,442	110,152,030	99,860,423	90.66	1.27	32.19	36.38	30.18
SF_4	100,082,932	96,724,472	84,749,793	87.62	0.54	29.84	40.47	29.16

Table 4.1: RNAseq summary metrics. UMI processing performed as described in Methods. UMI, Unique molecular indexes.

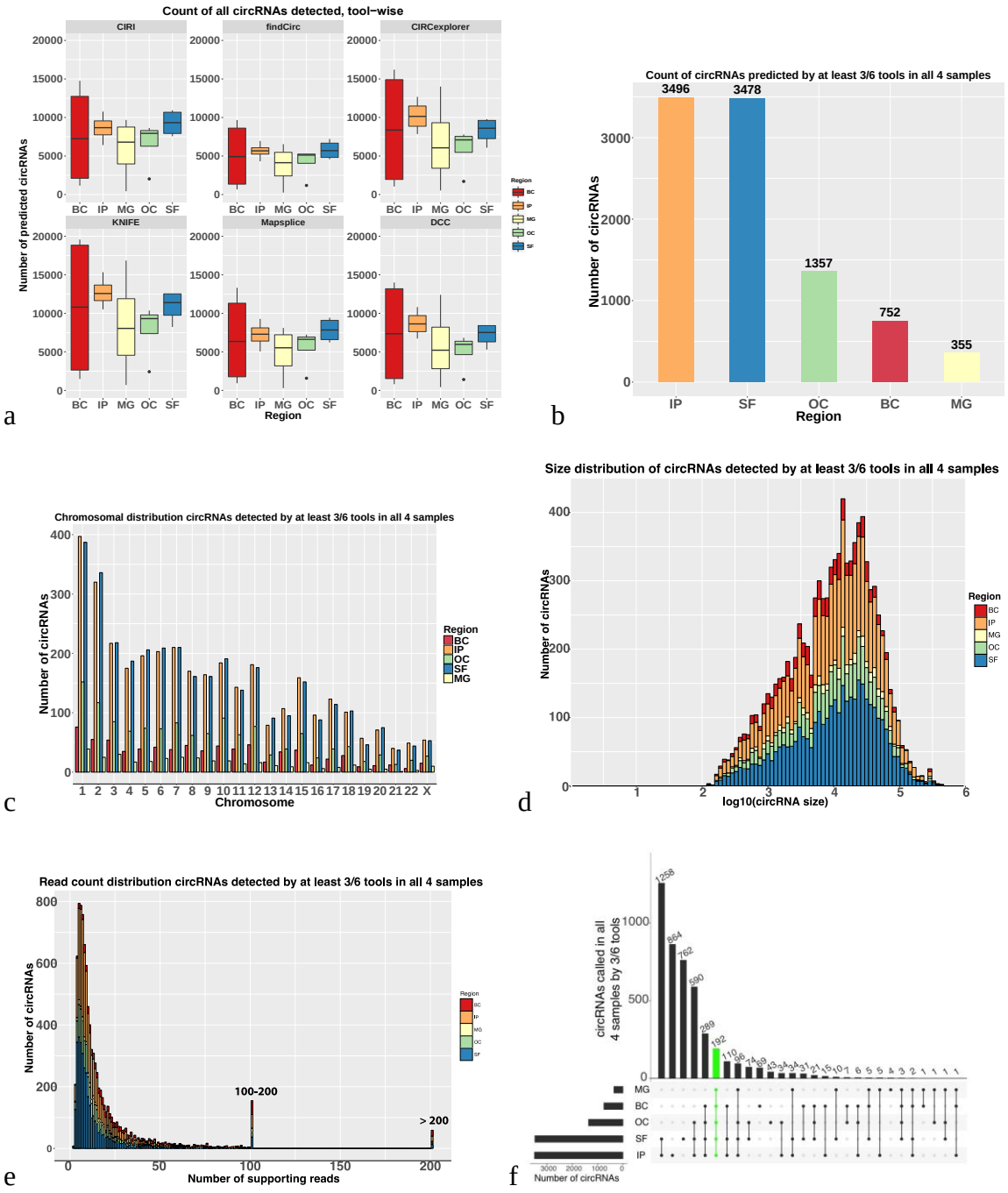


Figure 4.1: Summary of circRNA prediction results. (a) Tool-wise breakdown of circRNA detection across all brain regions; (b) Number of circRNAs called by at least three of six tools in all four samples; (c), (d), (e) Chromosomal, size and read count distribution of circRNAs called by at least three of six tools in all four samples (f) Intersection across regions of circRNAs called by at least three of six tools in all four samples; the green bar indicates the number of circRNAs called in all five regions. BC, cerebellum; IP, inferior parietal lobe; MG, middle temporal gyrus; OC, occipital cortex; SF, superior frontal gyrus

Region	CircRNA count	CDS	5'UTR	3'UTR	Intronic	Intergenic
BC	752	739	234	70	695	8
IP	3496	3418	1056	393	3290	31
MG	355	347	120	31	329	5
OC	1357	1332	412	144	1285	12
SF	3478	3410	1049	388	3296	23

Table 4.2: Summary of genomic region(s) of origin for high confidence circRNAs. CircRNA count is the number of circRNAs called by three of six tools in all four samples of that region.

We also observed in all brain regions that over 10% of the circRNAs map to chromosome 1, the longest chromosome, and <=1% circRNAs map to chromosomes 21 and 22 (Figure 4.1c). This is consistent with previous findings that the number of circRNAs detected is proportional to the length of the chromosome [26]. Further, the size distribution of all high confidence circRNAs spans a wide range between 129 and 432,649 bp (Figure 4.1d).

CircRNA abundance is quantified by the number of reads spanning the back-spliced junction. While 76% (7,189/9,438) of our list of high confidence circRNAs were supported by an average of only two to 20 reads, 214 circRNAs had an average of over 100 supporting reads, among which four were over 500 (Figure 4.1e). Further, to normalize for library size, we calculated the spliced reads per billion mapping (SRPBM) as described in [20] (Methods).

We next evaluated the extent of overlap among these high confidence circRNAs between the different brain regions (Figure 4.1f). We observed that 192 circRNAs were common across all five brain regions, while 69, 864, 4, 43 and 762 were unique to BC, IP, MG, OC and SF respectively. After accounting for these overlaps, there were 4,528

unique high confidence circRNAs across all the regions. The most abundant circRNAs that were common across all brain regions include 1:117,944,807-117,963,271, from gene *MAN1A2* (mannosidase alpha class 1A member 2), supported by an average SRPBM of 5,093 and 6:73,016,960-73,043,538, from gene *RIMS1* (regulating synaptic membrane exocytosis 1), supported by an average SRPBM of 4,631 across all the regions. Among the unique circRNAs, 16:31,230,415-31,230,840 from gene *TRIM72* (tripartite motif containing 72) in BC (Average SRPBM = 1100), 12:116,668,337-116,675,510 from gene *MED13L* (mediator complex subunit 13 like) in IP (Average SRPBM = 425.36), 3:170077486-170079217 from gene *SKIL* (SKI like proto-oncogene) in MG (Average SRPBM = 128.22), X:128250586-128256170 in OC (intergenic; Average SRPBM = 150.17) and 1:169947225-170001116 from gene *KIFAP3* (kinesin associated protein 3) in SF (Average SRPBM = 426.11) were the most abundant. Comparing our catalog of high confidence circRNAs with circBase, we observed that 67% (3,035/4,528) have been reported in at least one of the four studies deposited in the database; these studies evaluated various cell lines and tissue types, including cerebellum, diencephalon, SH-SY5Y cells, Hs68 cells, HeLa cells and HEK293 cells.

In order to compare circRNA abundance between different tissue types, we further sequenced 17 RNase R treated and rRNA depleted RNAseq libraries from the lung (LU, N= 3), liver (LV, N = 6), lymph node (LN, N = 4) and pancreas (PA, N = 4). We generated a median of 100,467,680 reads across all samples with an average mapping percentage of 84%. Applying the same analytical pipeline as above, we identified a total of 17,165 unique circRNAs across all samples, the majority of which were present in

only one of the samples in the respective tissue type. Comparing the number of circRNAs called by at least three of the six tools in each sample between our brain and other tissue type datasets, we observe a much higher abundance of brain circRNAs, as also reported in previous studies [20] (Figure 4.2). Upon filtering for circRNAs identified in all the samples by at least three of the six tools in a tissue type, 14, 5, 4 and 3 circRNAs were detected in LN, LU, PA and LV, respectively. Comparing these circRNAs with the high confidence circRNAs from the brain, we observe that 4,510 circRNAs were unique to the brain. Three circRNAs were present only in the lymph node- 7:117,825,700-117,828,459 from *LSM8* (LSM8 homolog, U6 small nuclear RNA associated), 7:50,358,643-50,367,353 from *IKZF1* (IKAROS family zinc finger 1) and 16:30,495,147-30,495,584 from *ITGAL* (integrin subunit alpha L) and one circRNA, 19:10,183,599-10,184,111 from *C3P1* (complement component 3 precursor pseudogene) was unique to the liver. Notably, previous studies have found abundant expression of the linear isoforms of these genes in the lymph node and liver [91, 92].

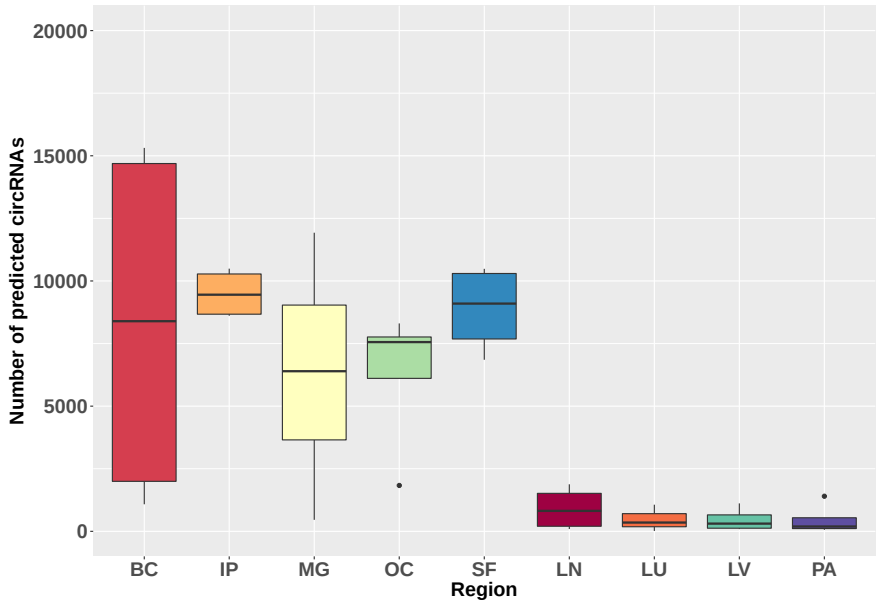


Figure 4.2: CircRNAs detected in brain vs. other tissue types. Other tissue types evaluated include LN, LU, LV and PA. LN, lymph node; LU, lung; LV, liver; PA, pancreas

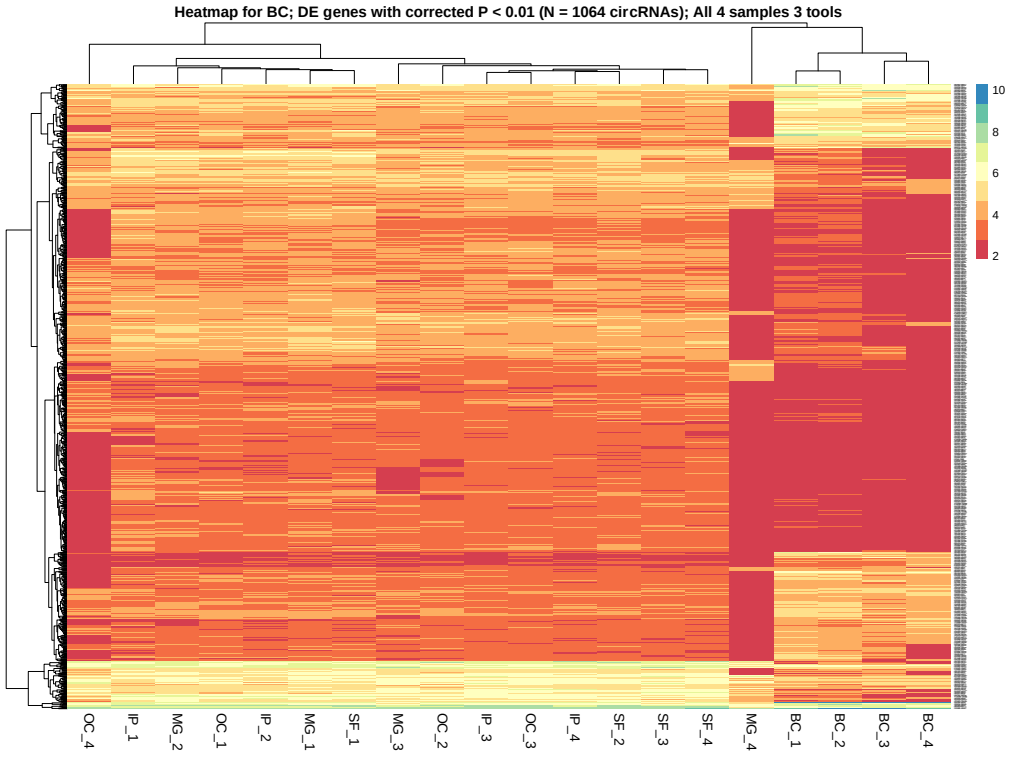
3.2 Region specific differential expression analysis and regulatory network prediction

Given the presence of abundant circRNAs in the brain, we asked if these brain circRNAs were expressed in a region specific manner. Using DESeq2, we identified differentially expressed (DE) circRNAs in each brain region compared to all other regions (Table 4.3, Figure 4.3). In the BC, 1,064 circRNAs were differentially expressed (corrected $P < 0.01$). 351 of these candidates also had a log2(fold change) > 3 or < -3. Fewer DE circRNAs were identified in the other four regions compared to BC (Table 4.3). We performed Ingenuity pathway analysis (IPA) on the genes from which these DE circRNAs arise. Our input genes were enriched in several neuronal and nervous system related functions and pathways commonly across all the regions. Examples include

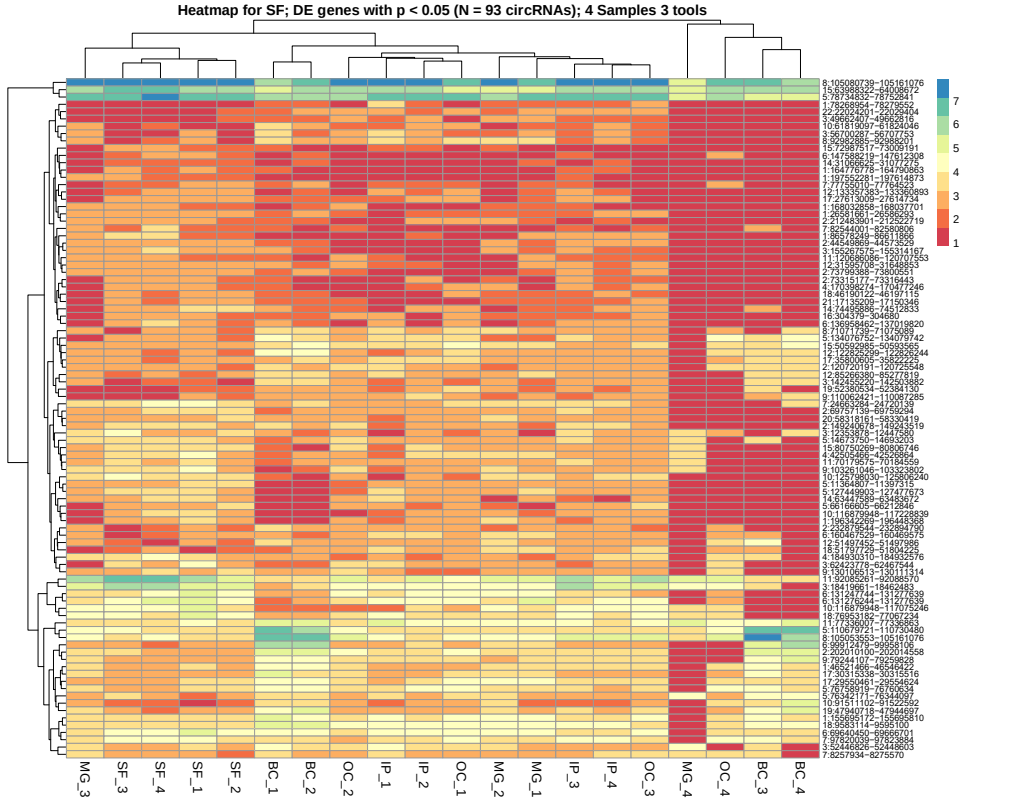
neuropathic pain signaling in dorsal horn neurons, neuritogenesis, synaptic transmission, morphology of the brain, excitatory post-synaptic potential and quantity of synaptic vesicles.

Region	# of DE circRNAs	# unique circRNA-miRNA interactions	# unique circRNAs	# unique miRNAs	# interactions with > 20 binding sites	# circRNAs in network	# miRNAs in network	# genes in network
BC	1064	13887	1053	1377	253	71	21	1142
SF	93	1377	92	524	14	7	5	312
MG	99	1431	98	517	11	11	6	393
OC	113	1244	110	374	28	2	2	65
IP	106	2054	106	635	29	10	10	334

Table 4.3: Summary of differentially expressed (DE) circRNAs and network analysis. DE cut off: uncorrected P < 0.05 for all regions except BC; corrected p-value < 0.01 for BC (N = 1,923 DE circRNAs with uncorrected P < 0.05). There were no circRNAs in IP, MG, OC and SF with corrected P < 0.01 or 0.05.

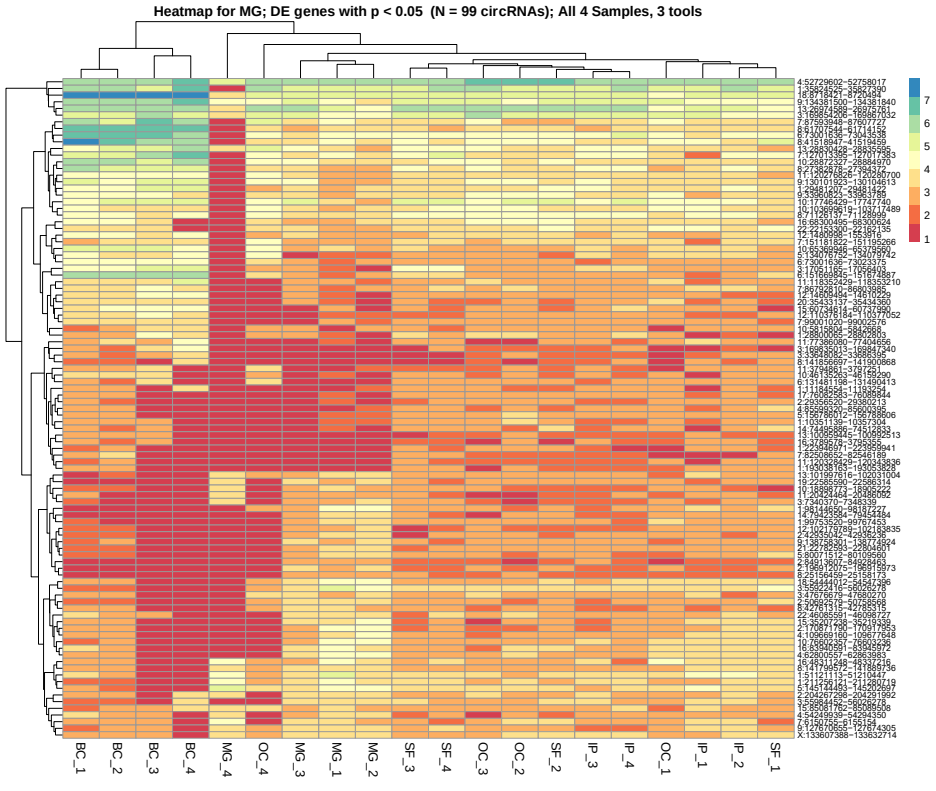


a.

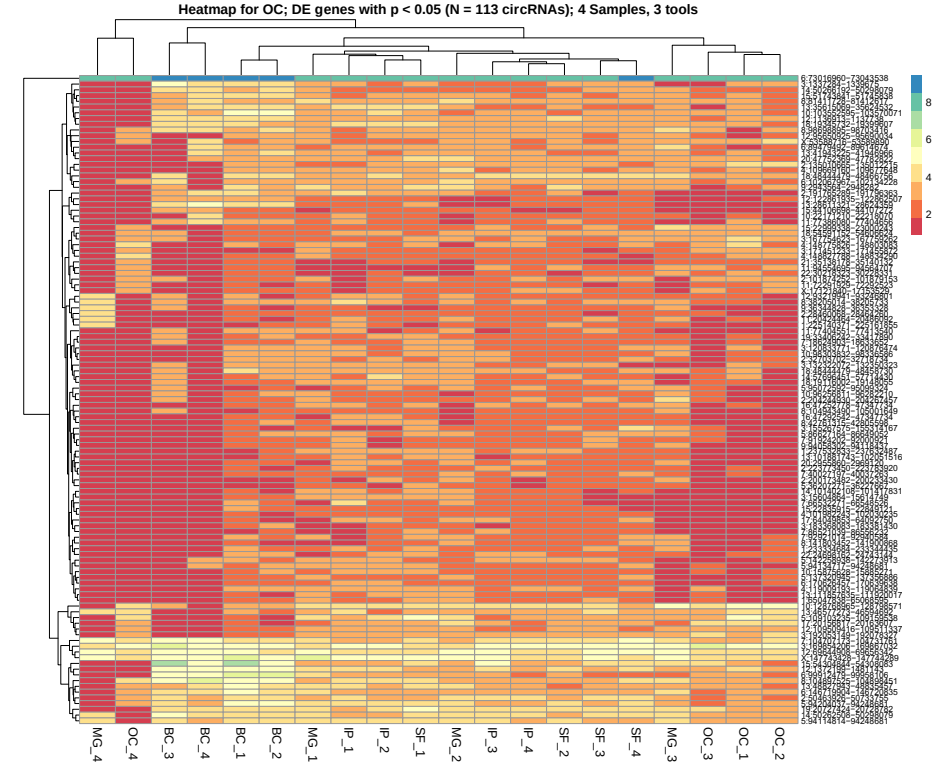


b.

Figure 4.3: Heatmaps of differentially expressed (DE) circRNAs (BC and SF maps shown)



c.



d.

Figure 4.3: Heatmaps of DE circRNAs (MG and OC maps shown)

e.

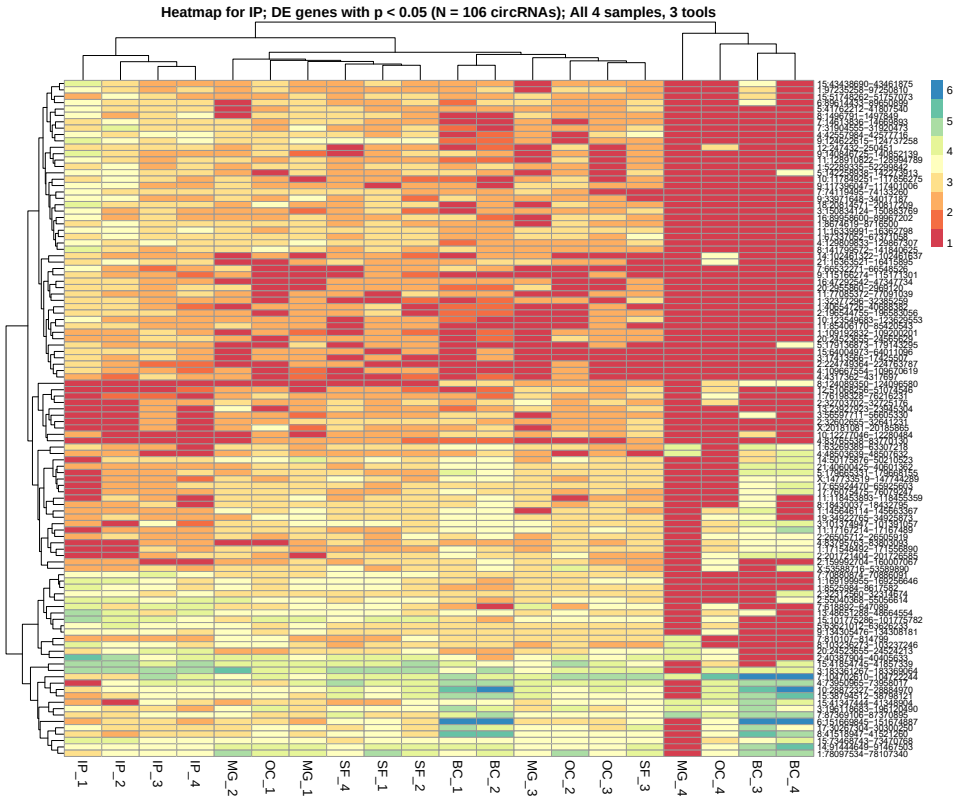


Figure 4.3: Heatmap of DE circRNAs (IP map shown)

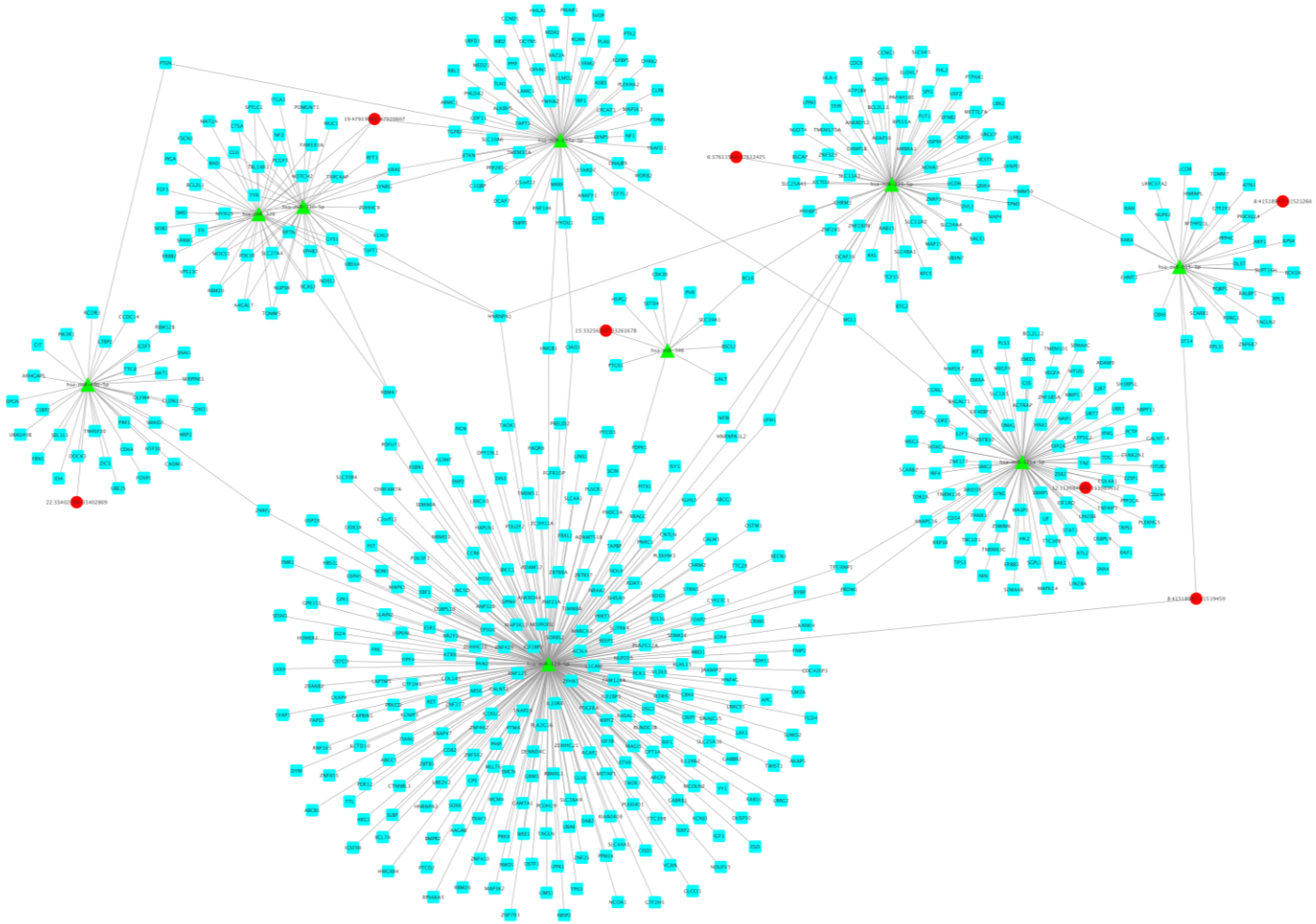
Using miRNA target prediction algorithms, miRNA targets of DE circRNAs in each brain region were predicted (Methods). Using a list of 2,588 published miRNAs from miRBase, we detected more than 1000 unique interactions between circRNAs and miRNAs predicted by both algorithms in each region (Table 4.3). The identified interactions represent binding sites for miRNAs on each circRNA candidate, predicted based on complementarity in the miRNA seed region (nucleotide positions 2-7 in the miRNA 5'-end). Interestingly, the unique interactions in each region were predicted to be between approximately 370 to 1370 unique miRNAs and approximately 90 to 1050 unique circRNAs in our input list, indicating that a single miRNA may bind several

different circRNAs and vice-versa. Further, among all the interactions detected, a small percentage was predicted by the miRanda algorithm to have 20 or more putative binding sites (Table 4.3).

MiRNAs are well known regulators of mRNAs and several databases have compiled predicted miRNA-mRNA interactions. We thus further employed the list of miRNA-mRNA interactions common in two databases: miRTarBase and TargetScan, to determine the target genes for our predicted miRNAs. Overall there were more than 300 unique target genes predicted for miRNAs in all regions except for the OC, which had 65 predicted target genes. By compiling all predicted miRNAs that may be regulated by identified circRNAs, as well as mRNAs that may be regulated by those miRNAs, we constructed circRNA-miRNA-mRNA regulatory networks for each brain region (Figure 4.4).

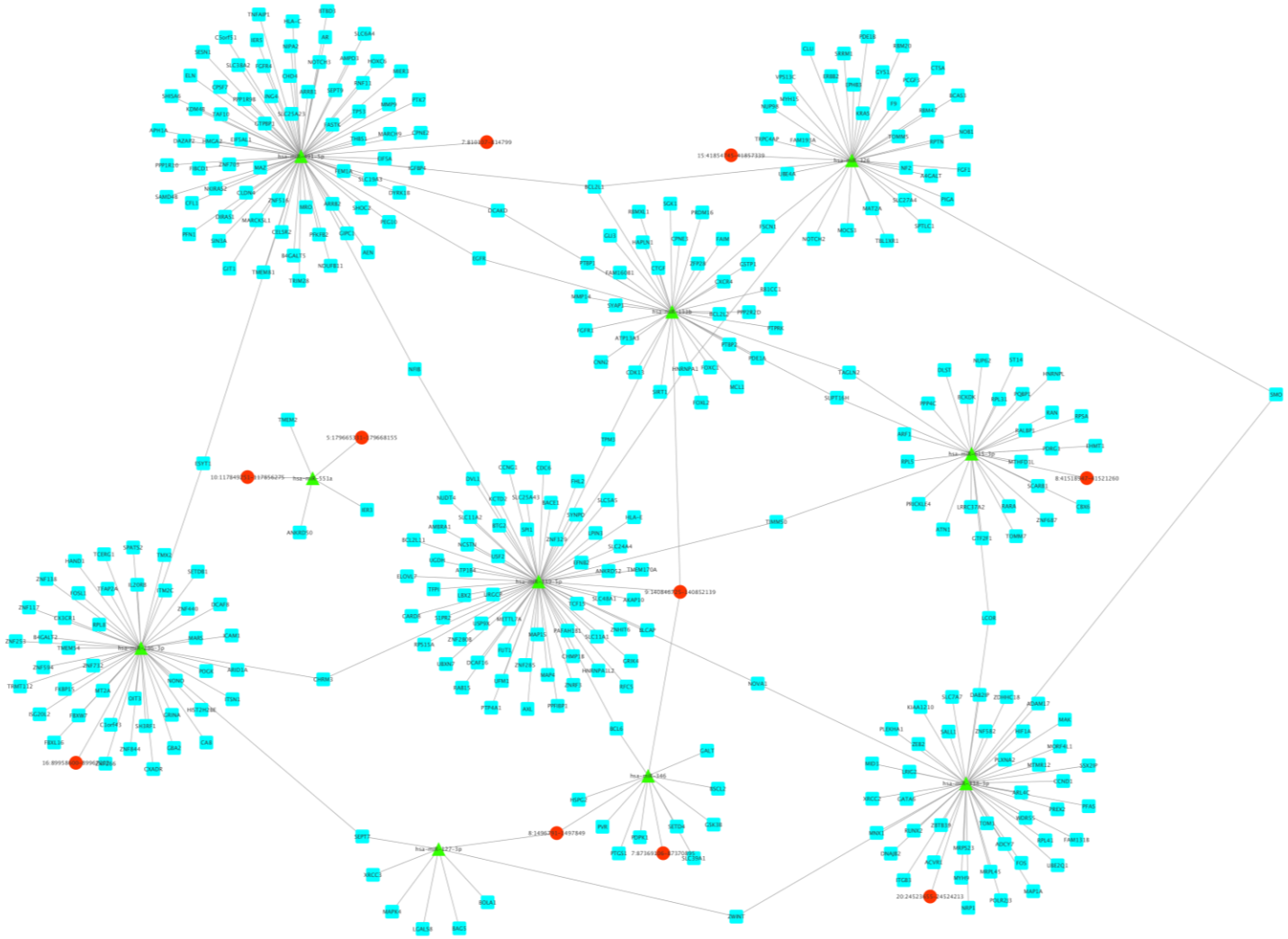
Lastly, we ran IPA to predict enriched pathways and functions among the predicted target genes in each region. Our analysis revealed significant enrichment (uncorrected p-value < 0.05) of several nervous system developmental functions, the majority of which were common across all the brain regions. Some of the common enriched functions include morphology of nervous system, brain size and formation, development, morphology and proliferation of neurons, synaptic plasticity, proliferation of neuroglia, Schwann cells and neuroblasts, and long-term potentiation. Genes associated with these functions include *NFIB* (nuclear factor I B), *PTEN* (phosphatase and tensin homolog), *NF1* (neurofibromin 1), *BACE1* (beta-secretase 1), *ITSN1*

(intersectin 1), *CHRM3* (cholinergic receptor muscarinic 3) and *AKAP5* (A kinase anchoring protein 5).



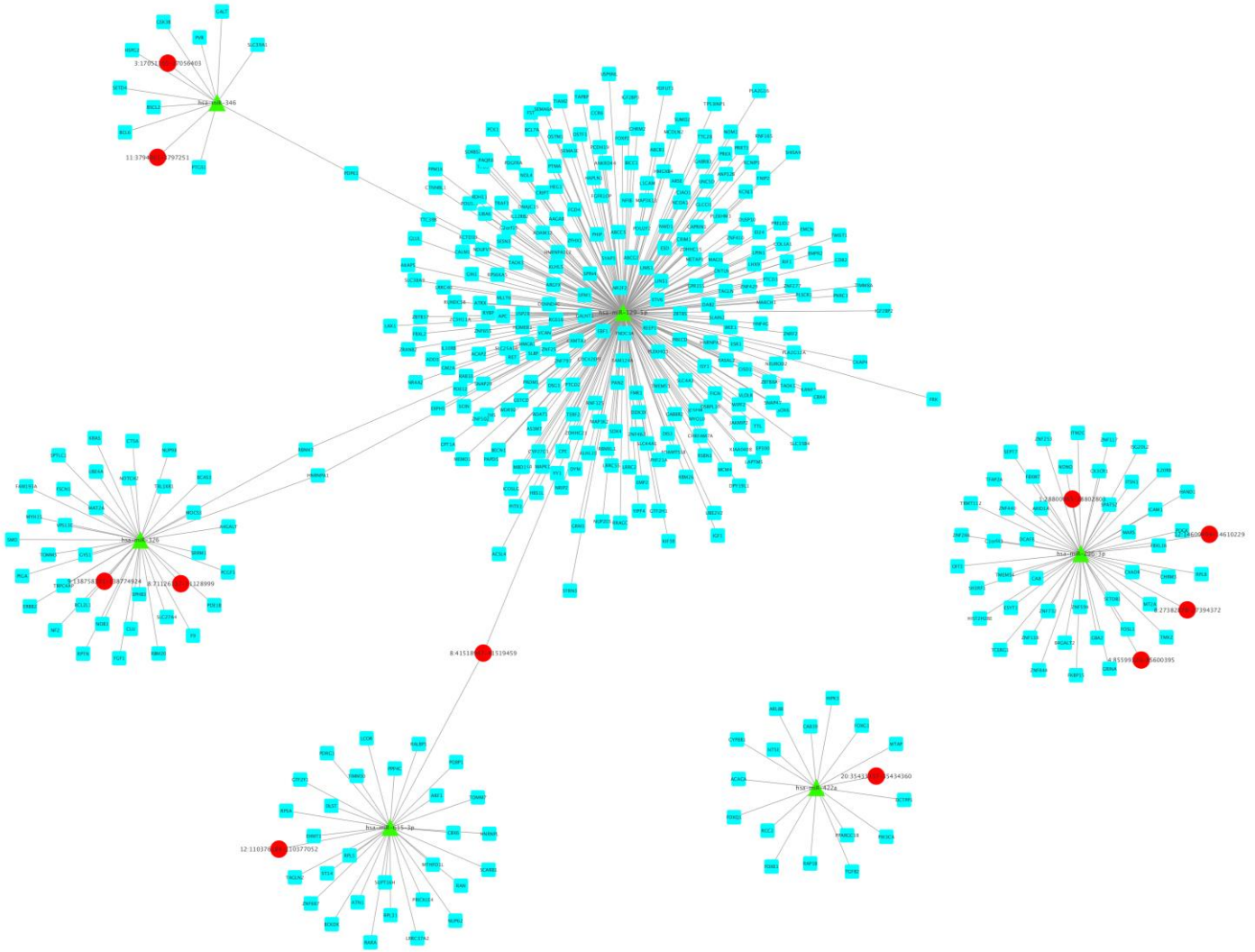
a.

Figure 4.4: Predicted circRNA-miRNA-mRNA regulatory network (BC)
Red circular nodes: circRNAs, green triangular nodes: miRNAs, blue square nodes: genes. mRNA, messenger RNA
(For the BC, only the top 50 up and downregulated circRNAs are depicted in the network for clearer visualization)



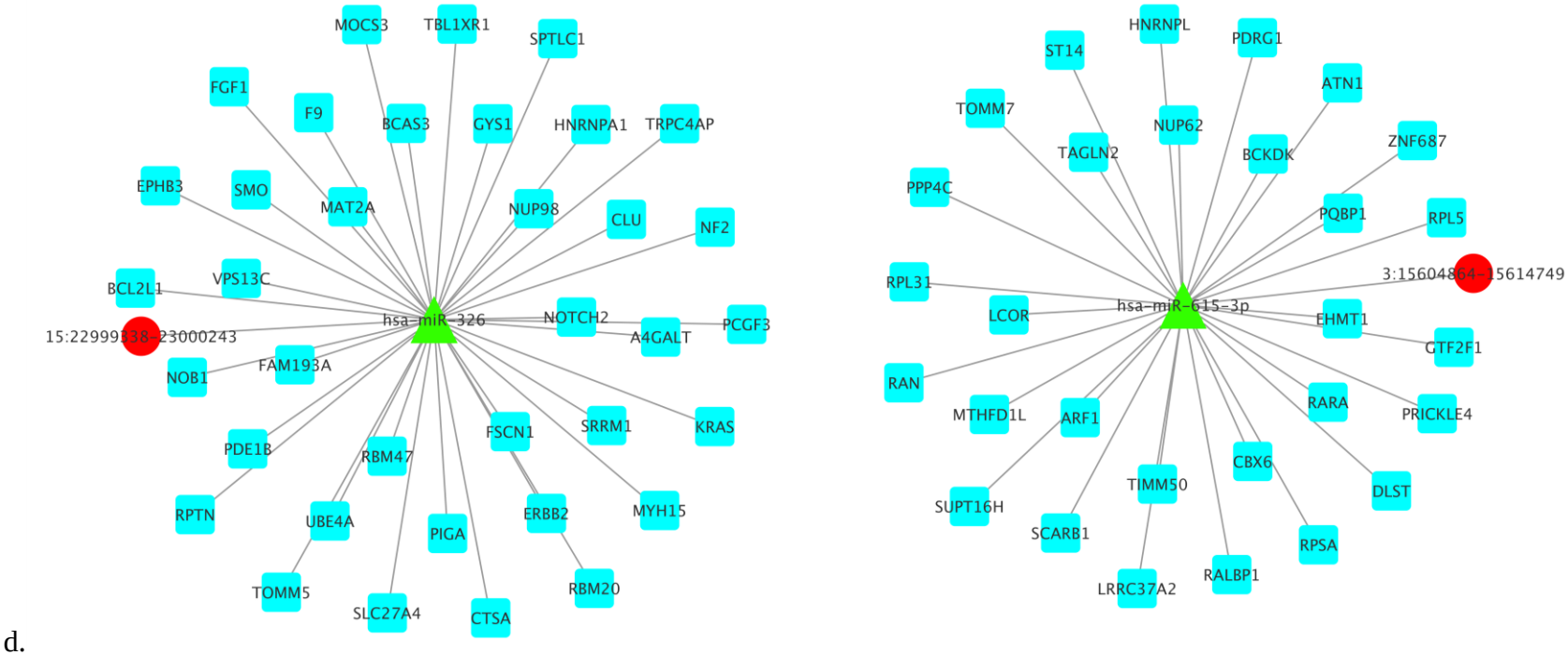
b.

Figure 4.4: Predicted circRNA-miRNA-mRNA regulatory network (IP) (All predicted miRNAs/mRNAs depicted)



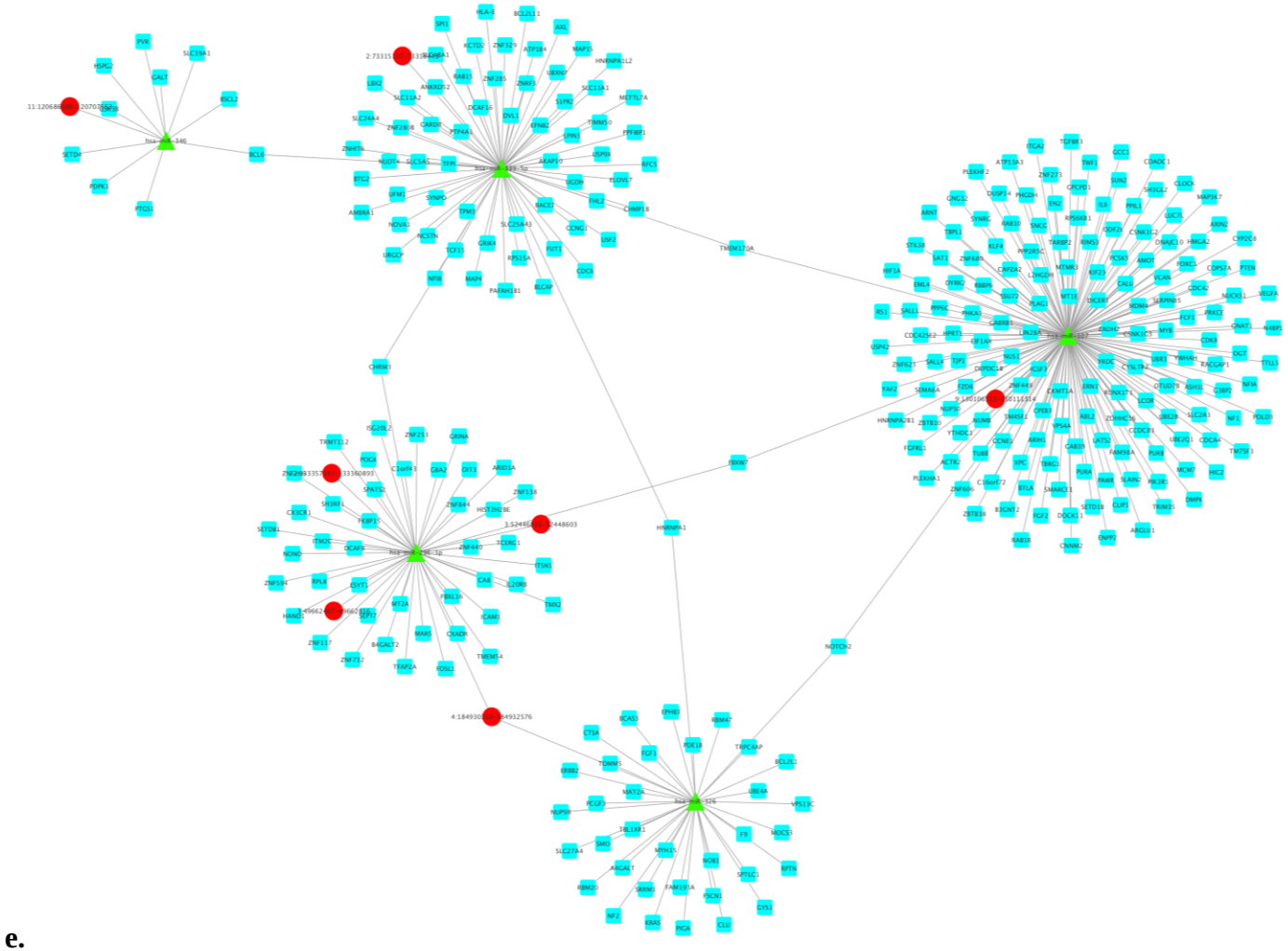
c.

Figure 4.4: Predicted circRNA-miRNA-mRNA regulatory network (MG) (All predicted miRNAs/mRNAs depicted)



d.

Figure 4.4: Predicted circRNA-miRNA-mRNA regulatory network (OC) (All predicted miRNAs/mRNAs depicted)



e.

Figure 4.4: Predicted circRNA-miRNA-mRNA regulatory network (SF) (All predicted miRNAs/mRNAs depicted)

Further, IPA also identified several pathways that were commonly enriched across the different brain regions (uncorrected $P < 0.05$; Appendix Tables 4.1 - 4.10). These common pathways include neuroinflammation signaling pathway associated with genes such as *TRAF3* (TNF receptor associated factor 3), *PIK3CA* (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha), *BDNF* (brain derived neurotrophic factor) and *AKT1* (AKT serine/threonine kinase 1), axonal guidance signaling associated with genes including *TUBB* (tubulin beta class I), *MAPK1* (mitogen-activated protein kinase 1), *DPYSL2* (dihydropyrimidinase like 2) and *GSK3B* (glycogen synthase kinase 3 beta), *NGF* (nerve growth factor) signaling associated with PI3Ks (phosphoinositide-3-kinases) - *PIK3CA* (PI3K, catalytic, alpha polypeptide), *PIK3R1* (PI3K, regulatory subunit 1 (alpha)), *MAPK1* and *MAP3K2* (MAP kinase kinase kinase 2), amyloid processing associated with *BACE1*, *GSK3B*, *MAPK1* and *CSNK2A1* (Casein Kinase 2 Alpha 1), *AMPK* signaling associated with *ARID1A* (AT-rich interaction domain 1A), *AK2* (adenylate kinase 2), *CHRM3 and RAB6A* (RAB6A, member RAS oncogene family), and neuregulin signaling associated with *PTEN*, *NRG1* (Neuregulin 1), kinases *MAP2K1* (MAP kinase kinase 1), *MAPK1 and PRKCE* (protein kinase C epsilon).

IPA on each region also predicted a few functions and pathways that were unique to that region. The BC is posterior to the brain stem, and receives information from the sensory systems, spinal cord, and other parts of the brain to regulate motor movements. One of the pathways identified uniquely in BC was the NF-kB (nuclear factor kappa B) signaling pathway (uncorrected P=0.00003). This is consistent with previous findings that identified constitutive NF-kB activity in rodent brain regions including the

cerebellum [93]. NF-kB is required for cerebellar granule neurons survival [94] as well as plays crucial roles in the nervous system such as long term response to synaptic plasticity, survival and apoptosis of neurons [95, 96]. 26 genes from our input list were associated with this pathway and include the PI3Ks *PIK3CA* and *PIK3R1,* which are responsible for coordinating cell functions such as proliferation and survival, MAPK genes such as *MAP3K7* (MAP kinase kinase kinase 7), *MAP2K7* (MAP kinase kinase 7) and *MAP3K1* (MAP kinase kinase 1). Other genes in this pathway include *AKT1, RELA* (RELA proto-oncogene, NF-kB subunit), *GSK3B* and *TRAF6* (TNF receptor associated factor 6). Other enriched pathways include IL-3 (Interleukin-3) and IL-6 (Interleukin-6) signaling, neuropathic pain signaling in dorsal horn neurons and reelin signaling in neurons (Appendix Tables 4.1, 4.2).

The SF is located at the superior aspect of the prefrontal cortex and constitutes about one-third of the frontal lobe in the human brain. It is involved in executive functions including self-awareness, working memory, and attention [97]. Among the DE circRNAs in this region, the ones mapping to *PLCH1* (Phospholipase C Eta 1), *AAK1* (AP2 Associated Kinase 1) and *RIMS2* (Regulating Synaptic Membrane Exocytosis 2) had the highest fold changes. IPA on the 312 genes in the circRNA-miRNA-mRNA network identified 92 enriched pathways and 217 nervous system developmental functions (uncorrected $P < 0.05$). Among the enriched pathways, the ones unique to this region include the Wnt/Ca+ pathway, actin nucleation by ARP-WASP (actin related protein-Wiskott-Aldrich syndrome protein) complex and cell cycle control of chromosomal replication. Wnt/Ca^{2+} and Wnt/β-catenin pathways participate in pre- and

post-synaptic receptor localization as well as neuronal pathfinding [98]. Genes associated with this pathway include *DVL2* (Dishevelled Segment Polarity Protein 2), *FZD6* (Frizzled Class Receptor 6), *SMO* (Smoothened, Frizzled Class Receptor) and *GSK3B*. While functions such as formation of the forebrain, branching of neurites, axonogenesis, long term potentiation of hippocampal CA1 region were enriched in SF as well as other brain regions, a few functions were also uniquely enriched in the SF, including proliferation of neural precursor cells, outgrowth of axons, formation of dendritic spines, and migration of neuroglia (Appendix Tables 4.3, 4.4).

The MG is located in the temporal lobe between the superior and inferior temporal gyri and is involved in face recognition, ascertaining distance, memory and emotion. DESeq2 analysis revealed that circRNAs from *DNAH7* (dynein axonemal heavy chain 7), *NSUN6* (NOP2/Sun RNA methyltransferase family member 6), *EPHX2* (epoxide hydrolase 2), and *AKAP7* (A-Kinase anchoring protein 7) had the highest fold changes. IPA analysis on genes from the circRNA-miRNA-mRNA network identified 86 significantly enriched pathways and 164 nervous system developmental functions (uncorrected p-value < 0.05) (Appendix Tables 4.5, 4.6). Among these, synaptic long term potentiation, glutamate receptor signaling and glutamine biosynthesis were uniquely identified in the MG. Long term potentiation is one of the major cellular mechanisms underlying synaptic plasticity as well as learning and memory. A few other genes associated with this pathway from our input list include *CALM1* (calmodulin 1), which encodes for a calcium binding protein, *GRM3* (Glutamate Metabotropic Receptor 3) which encodes G protein coupled receptor proteins and is involved in most aspects of

normal brain functions, *EP300* (E1A Binding Protein P300) which encodes the adenovirus E1A-associated cellular p300 and *PRKCD* (Protein Kinase C Delta) whose encoded protein is a positive regulator of cell cycle progression. Further, the neurodevelopmental function differentiation of oligodendrocyte precursor cells were also identified uniquely in MG.

The OC is located at the dorsal aspect of the brain, behind the temporal and parietal lobes, and is primarily responsible for processing visual information. Among the DE circRNAs in this region, 81% were downregulated compared to all other regions. IPA analysis on genes from the circRNA-miRNA-mRNA network identified significant enrichment in 45 pathways and 52 nervous system developmental functions (uncorrected $P < 0.05$) (Appendix Tables 4.7, 4.8). This is relatively lower compared to other regions due to the smaller number of genes identified in the network analysis. Some of the functions identified in this region that were also present in other regions include myelination, synaptic plasticity, proliferation of neurons and activation of neuroglia, associated with genes including *ATN1* (Atrophin 1), *NF2* (Neurofibromin 2), *CLU* (Clusterin), *BCKDK* (Branched Chain Ketoacid Dehydrogenase Kinase) and *FGF1* (Fibroblast Growth Factor 1). Apart from the common functions and pathways identified in this region, a few unique pathways such as glycogen biosynthesis, histidine degradation, and folate transformations were also identified.

The IP lies below the horizontal portion of the intraparietal sulcus and is involved in the perception of emotions in facial stimuli, interpretation of sensory information and language and mathematical operations. 63 of the 106 DE circRNAs in this region

demonstrated a positive fold change (uncorrected $P < 0.05$). IPA on the 334 genes from the network revealed significant enrichment of 114 pathways and 162 neurodevelopmental functions (uncorrected $P < 0.05$) (Appendix Tables 4.9, 4.10). Several of the identified pathways and functions were also identified in other regions including neuroinflammation signaling pathway, axonal guidance signaling, and IL-7 signaling pathways. Further, pathways including sonic hedgehog signaling, OX40 signaling pathway and the role of p14/p19ARF in tumor suppression were uniquely identified in IP.

3.3 Assembly based detection of circRNAs

We next applied DeFuCir, the assembly based detection approach described in Chapter 3 on these samples. Our de novo approach detected 10,000 to 67,000 circRNA contigs across the samples, with a median of 25,860 (Table 4.4). Compared with candidates identified by six existing circRNA detection tools and allowing a 10 bp window for the match (Methods), we observed only about 6% overlap between the two approaches. However, for these 6% of overlapping candidates, the full-length sequences of the circRNA transcripts are constructed using trinity assembly. We hence report these overlapping candidates as our list of high-confidence full length circRNAs.

Among these overlapping circRNA candidates, 6:73005639-73043538 in the BC from gene *RIMS1* (SRPBM = 30037.4), 6:54013853-54067031 in the IP from gene *MLIP* (Muscular LMNA Interacting Protein; SRPBM = 3992.89), 6:73016960-73043538 in the SF and OC from *RIMS1* (SRPBM = 3712.74 in SF, 3703.26 in OC) and 8:105080739-

105161076 in the MG from *RIMS2* (SRPBM = 2488.05) demonstrated the highest expression in terms of SRPBMs.

Sample ID	Total number of contigs from trinity	# of contigs with back-spliced soft clipping signals (circRNAs)	Overlap with all candidates from existing approaches	Overlap with candidates called by 3/6 tools from regular approach
BC_1	715,292	18,667	2265	1544
BC_2	1,015,381	30,785	2195	1576
BC_3	1,342,745	52,276	597	422
BC_4	1,260,320	67,798	309	197
IP_1	409,760	10,844	1573	945
IP_2	575,742	15,150	1723	1111
IP_3	926,497	29,065	1348	920
IP_4	885,136	30,235	1292	915
MG_1	585,278	15,508	1745	1083
MG_2	685,177	17,787	1130	751
MG_3	1,288,266	41,132	836	570
MG_4	1,267,560	46,107	129	85
OC_1	720,260	17,308	1157	802
OC_2	782,643	18,330	1062	737
OC_3	768,926	22,655	1016	697
OC_4	1,491,938	65,442	424	279
SF_1	642,558	16,434	1380	975
SF_2	835,582	20,501	1319	917
SF_3	1,101,862	33,218	1015	719
SF_4	917,800	34,167	1266	864

Table 4.4: Summary of DeFuCir results

3.4 Assembly based *in silico* validation of circRNAs

In order to *in silico* validate selected circRNAs, we next implemented CircValidator, the assembly based in silico validation approach described in Chapter 3. Among the differentially expressed circRNAs, we observed that on an average, 41% of the candidates validated with the high-stringency threshold while 46% and 47% validated

with the low-stringency and very-low stringency thresholds (Table 4.5a). On the other hand, when we applied CircValidator on candidates called by all tools in all samples, we observed a higher validation rate of 80% using the high-stringency threshold and 86% using the low-stringency threshold (Table 4.5b). These results demonstrate that an ensemble approach yields higher confidence circRNA results compared to using a single, or a few, tools, as previously reported (Hansen 2015).

a)

Sample ID	# Input for validation	# validated HS	# validated MS	# validated LS	# validated VLS
BC_1	100	47	47	47	49
BC_2	100	42	45	47	48
BC_3	100	36	39	39	39
BC_4	100	33	37	37	39
IP_1	106	55	63	64	64
IP_2	106	46	54	56	57
IP_3	106	45	52	54	54
IP_4	106	47	54	54	54
MG_1	99	48	50	50	51
MG_2	99	49	51	53	54
MG_3	99	41	42	43	44
MG_4	99	10	11	11	11
OC_1	113	42	46	46	47
OC_2	113	36	37	37	37
OC_3	113	42	48	48	48
OC_4	113	18	20	21	22
SF_1	93	52	59	60	62
SF_2	93	52	58	59	61
SF_3	93	44	49	50	52
SF_4	93	50	54	55	56

Table 4.5: Summary of ACValidator results when run on a) DE circRNAs from each region

b)

Sample ID	# Input for validation	# validated HS	# validated MS	# validated LS	# validated VLS
BC_1	50	48	48	48	48
BC_2	50	44	45	45	45
BC_3	50	39	42	43	43
BC_4	50	35	40	41	42
IP_1	285	219	231	234	237
IP_2	285	225	241	243	247
IP_3	285	207	222	224	226
IP_4	285	221	233	234	237
MG_1	12	12	12	12	12
MG_2	12	12	12	12	12
MG_3	12	9	10	10	10
MG_4	12	5	7	7	7
OC_1	97	75	81	81	81
OC_2	97	79	85	85	86
OC_3	97	80	87	87	87
OC_4	97	73	79	80	82
SF_1	299	225	239	241	243
SF_2	299	218	236	240	241
SF_3	299	217	230	234	234
SF_4	299	232	245	246	246

Table 4.5: Summary of ACValidator results when run on b) circRNAs called by all tools in all samples in a region. HS, high stringency cut off; MS, medium stringency cut off; LS, low stringency cut off; VLS, very low stringency cut off

4. Conclusions

In this study, we profiled circRNA expression in five brain regions from four healthy aged individuals and predicted their impact on transcriptional regulatory networks in each region. Based on the results from ensemble circRNA detection analysis using six tools, we identified a total of 4,528 circRNAs that were called in all samples across brain regions by at least three of the six tools. While 192 circRNAs were common across all the regions, each region also had unique circRNAs. Some of these include

circRNAs mapping to *MED13L, TRIM72, SKIL* and *KIFAP3.* Among the circRNAs detected across all regions, *RIMS1* had a high SRPBM of 4,631. The protein encoded by this gene regulates synaptic vesicle exocytosis and is associated with cone-rod dystrophy and retinitis pigmentosa. Recently, Rybak Wolf et al [20] also reported abundant expression of *RIMS2* in their mammalian brain circRNA study.

MiRNA target prediction analyses on differentially expressed circRNAs revealed 1,244 to 13,887 unique circRNA-miRNA interactions across brain regions. Furthermore, 65 to 1,142 target genes were identified for these miRNAs. IPA on the genes in each region's network revealed significant enrichment of several neurodevelopmental and nervous system related functions and pathways. While the majority of these enriched ontologies were common across all regions, a few pathways and functions were also unique to each region.

Using the assembly analysis workflows described in Chapter 3, we identified a list of high-confidence full length circRNAs, as well as in silico validated approximately 45% of DE circRNAs and more than 80% of circRNAs called by all tools in all samples. Notably, the assembly based de novo detection approach identified *RIMS1* and *RIMS2* in four of the five brain regions with SRPBMs > 2000. Interestingly, a different isoform of *RIMS1* was detected in the BC compared to those identified in the SF and OC.

Although our study lends new insight into region specific circRNA expression and regulatory networks, we are limited by a few technical caveats. Primarily, while RNase R treatment reduces the population of linear RNAs in a sample, it is not well understood if this depletion step introduces any biases in circRNA detection and if this

step has the potential to deplete circRNAs. Previous studies have reported that in some cases, circRNAs are sensitive to RNase R [2, 30, 42]. In order to understand whether the detected circRNA expression profiles are specific to healthy aged brains, deeper investigations using healthy younger controls as well as diseased brain samples will be needed. Similarly, additional cell-specific characterization of circRNAs using circRNA-enriched datasets will help us understand whether their expression varies across cell types. Such analyses on the linear transcriptome in past have helped identify cell-specific contributions to neurological diseases such as in AD [64, 99, 100].

Overall, our study describes circRNA expression in a brain region-specific manner using circRNA enriched RNAseq data and thus contributes to our understanding of this more newly classified class of RNAs. Through our above analyses, we have established a valuable reference dataset of their expression profiles in healthy elderly individuals for the circRNA research community. This resource, along with existing databases such as circBase and future experimental studies, will help us to better understand how and whether circRNAs represent a novel layer of transcriptional regulation in the brain as well as pave an avenue towards evaluating their utility in biomedical applications.

5. Methods

5.1 Sample acquisition

Postmortem brain samples were collected at the Banner Sun Health Research Institute's Brain and Body Donation Program from five functionally distinct brain

regions from six clinically normal aged subjects (protocols as described in Chapter 2). Total RNA extracted from the pancreas, liver, lung and lymph node tissues of four healthy subjects in each tissue type was purchased from Proteogenex (Culver City, CA).

5.2 RNA isolation and RNase R treatment

For the brain samples, total RNA was extracted from 100 mg of each of the brain regions. Tissue was divided on dry ice into 50 mg pieces and placed in 2 mL Rino tubes (NextAdvance, Troy, NY) containing an equal amount of Rino beads (NextAdvance, Troy, NY). The tissue was homogenized by adding 300 µl of lysis buffer and then using a Bullet Blender (NextAdvance, Troy, NY) at speed 8 for 3 mins followed by a QIAshredder (Qiagen, Hilden, Germany) column that was spun for 2 mins at full speed. The homogenate was processed according to the mirVana miRNA isolation kit protocol (Thermo Fisher, Waltham, MA). The 100 µl of eluted RNA was taken into a RNA Cleanup and Concentrator-5 (Zymo Research, Irvine, CA), which consisted of an on-column DNase treatment. RNA quality and quantity was measured with a NanoDrop 1000 (Thermo Fisher, Waltham, MA) and a RNA screen tape on the 4200 Tapestation (Agilent Technologies, Santa Clara, CA) according to manufacturer's protocols.

For the RNase R treatment of both brain and other tissue type samples, RNA was normalized to 102.6 ng/µl (4µg of RNA in 39 ul) using nuclease-free water. In a 1.5 mL DNA Lo-Bind tube (Eppendorf, Hamburg, Germany), 5 µl of 10X RNase R reaction buffer and 6 µl of RNase R 2U/µl (Epicentre, Madison, WI) were added to the 39 µl of RNA. The reaction mix was pipette-mixed 10 times and then placed in a water bath for

10 mins at 37 °C. After incubation, the treated RNA was placed on ice and immediately cleaned up with the Cleanup Concentrator-5, this time without DNase treatment. The RNA was eluted with water and all 10 µl were transferred into a 96-well plate for library preparation. The sealed plate was frozen at -80 °C for up to five days before library preparation was started.

5.3 Library preparation and paired-end sequencing

Sequencing libraries were prepared using the Illumina TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat Kit (Illumina, San Diego, CA) following the manufacturer's protocol except where noted. For the incorporation of unique molecular indexes (UMIs), 5ul of 500 nM xGen Dual Index UMI Adapters (IDT, Skokie, IL) were used in place of the TruSeq adapters and additional re-suspension buffer. For DNA fragment enrichment, 8 cycles (brain samples) and 15 cycles (other organs) of PCR were performed. Completed libraries were quantified using the High Sensitivity DNA Screen tape and reagents on the 4200 Tapestation and with the Qubit dsDNA High Sensitivity Assay kit (Thermo Fisher, Waltham, MA). Equimolar amounts of library were pooled and sequenced on the HiSeq (Illumina, San Diego, CA) to generate paired 82 bp reads.

5.4 Data analysis

Raw sequencing data in the form of basecall files (BCLs) were converted to FASTQ format using Illumina's bcl2fastq conversion software and quality checked using fastQC [80]. FASTQs were quality trimmed to 76 bp across all samples using fastx_trimmer.

Since UMIs were incorporated in our samples, we first ran bwa mem on the trimmed FASTQs followed by UMI aware dup marking using Picard. These bams were then re-converted to FASTQ using bedtools [83] and were then run through six different circRNA prediction algorithms - CIRCexplorer (v1.1.10) [42], CIRI (v2) [46], DCC (v0.4.4) [44], find_circ (v1) [3], Mapsplice (v2.1.8) [36] and KNIFE (v1.4) [45] using default parameter settings. Samples from two of the six donors were dropped from further analysis due to outliers in circRNA detection. CircRNAs from each sample with at least two supporting reads were used for further downstream processing and analyses. CIRI, Mapsplice and DCC produce 1-based circRNA co-ordinates, and were therefore converted to 0-based co-ordinates to be consistent with the other three algorithms.

5.4.1 CircRNA normalization

CircRNA reads were normalized as described in Chapter 2, using the formula

$$SRPBM = \frac{Number\ of\ back-spliced\ reads}{Total\ number\ of\ mapped\ reads} \times\ 1{,}000{,}000{,}000$$

5.4.2 Differential expression analysis

DESeq2 analysis was performed on circRNA candidates detected in all four samples by at least three of the six tools, to identify DE circRNAs in each brain region compared to all other regions. We filtered for DE circRNAs with Benjamini and Hochberg (BH) corrected $P < 0.01$ for the BC, uncorrected $P < 0.05$ for the other regions, and generated expression heatmaps for the DE circRNAs.

5.4.3 miRNA target prediction

We next conducted miRNA binding site prediction using the miRanda [65] and RNAHybrid [66] algorithms on DE circRNAs with corrected $P < 0.01$ in the BC and

uncorrected $P < 0.05$ in all other regions (since the IP, OC, MG and SF did not have any circRNAs with corrected $P < 0.01$). MiRanda and RNAHybrid algorithms are briefly summarized in Chapter 2, Methods. Only those circRNA-miRNA interactions predicted by both the algorithms were used for downstream network construction and analyses. We also required the interactions to have RNAHybrid predicted minimum free energy (MFE) < -20 and uncorrected $P < 0.05$, as well as a miRanda match score >= 150.

5.4.4 CircRNA-miRNA-mRNA network construction

MiRNA-mRNA interactions common to both miRTarBase [69] and TargetScan [70] were used to determine the gene targets of each filtered miRNA from the miRNA target prediction analysis. Using these circRNA-miRNA and miRNA-mRNA interaction predictions, we outlined circRNA-miRNA-mRNA regulatory networks for each brain region using custom python scripts and visualized the networks using Cytoscape [68].

5.4.5 Pathway analysis

Using the list of mRNAs identified in the circRNA-miRNA-mRNA networks, we performed enrichment analysis using IPA v01-12 from Qiagen to predict neurodevelopmental functions and pathways enriched in the different brain regions. The results were filtered for functions and pathways with an uncorrected $P < 0.05$.

5.4.6 Assembly based de novo detection of full length circRNAs

We applied DeFuCir, the assembly based de novo detection workflow described in Chapter 3 in order to detect full length circRNA transcripts. Briefly, all UMI dup removed FASTQs were assembled using Trinity, followed by aligning the trinity contigs to GRCh37 reference genome. Custom python scripts were then run to detect contigs that

had evidence of back-splicing. We further compared circRNAs detected using the assembly approach with those from existing circRNA detection tools allowing a 10 bp window for the match. Specifically, we compared them against the union of circRNAs detected by all tools as well as circRNAs called by at least three of six tools. The overlapping candidates constitute our list of high confidence full length circRNAs.

5.4.7 *In silico* validation of circRNAs

ACValidator, described in Chapter 3, was run to *in silico* validate selected circRNA candidates. Briefly, given a circRNA coordinate and the input alignment file, circValidator extracts reads from a fixed window on either side of the circRNA junction and assembles those reads. The assembled contigs are then interrogated for overlap with the circRNA junction sequence, with different stringency cut-offs of the number of overlapping bases (high stringency - 30 bp, medium stringency - 20 bp, low stringency - 10 bp and very low stringency - 5 bp overlap on either side of the junction to validate). Using this approach, we validated the list of differentially expressed circRNAs detected in each region, as well as those circRNAs called by all six tools in all four samples in a region.

6. Summary

In summary, we profiled circRNA expression in a brain region specific manner and identified circRNA candidates unique to each of the five brain regions analyzed, as well as common across the regions. Further, we detected differentially expressed circRNAs in each brain region compared to all other regions using DESeq2. MiRNA target prediction analysis and subsequent identification of circRNA-miRNA-mRNA networks was then performed on the differentially expressed circRNAs (corrected $P < 0.01$ for BC, uncorrected $P < 0.05$ for all other regions) from each region. Enrichment analysis on the genes from each region's regulatory network identified several nervous system related functions and pathways that were commonly impacted as well as unique to each brain region. Lastly, using ACValidator, we *in silico* validate 45% of DE circRNAs as well as over 80% of the ones called by all tools in all the samples.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

1. Conclusions

CircRNAs represent a newer area of research in transcriptomics and are a class of endogenous, non-coding RNAs that are formed when exons back-splice to each other. Since 2012, multiple circRNA studies have reported that they are pervasively expressed in eukaryotes, especially in the mammalian brain. While the functional role and impact of circRNAs remains to be clarified, they have been found to regulate miRNAs as well as parental gene transcription and might thus have key roles in transcriptional regulation. Although circRNAs have continued to gain attention, our understanding of their expression in a cell-, tissue- and brain region-specific context remains limited. Further, computational algorithms produce varied results in terms of the detected circRNAs. This thesis has attempted to address some of these gaps in circRNA research.

We first used existing circRNA detection algorithms and identified circRNAs in RNAseq data previously generated from PC astrocytes microdissected from AD patients (N=10) and healthy elderly controls (N=10). In subsequent analyses, we predicted miRNA targets of circRNAs, and constructed regulatory circRNA-miRNA-mRNA networks using in silico methods. Notably, the widely reported CDR1as circRNA was predicted to have binding sites for 74 distinct miRNAs and 63 binding sites for miR-7. Pathway analysis of the genes regulated by these miRNAs identified significantly enriched immune response pathways, which is consistent with the known function of

astrocytes as immune sensors in the brain. While we did not detect recurrent differentially expressed circRNAs in the context of healthy
controls or AD, we report for the first time circRNAs and their potential regulatory impact in a cell-specific and region-specific manner in aged subjects. These predicted regulatory network and pathway analyses may help provide new insights into transcriptional regulation in the brain.

Although available circRNA computational tools can identify circRNAs in RNAseq data, these tools do not capture the full length sequence of the circRNA and further, produce divergent results. We developed two novel bioinformatics approaches, DeFuCir and ACValidator to address these challenges. DeFuCir is an assembly based workflow for the characterization of full length circRNAs and that leverages soft-clipping signals in an alignment file. Though DeFuCir detects a smaller number of candidates, it achieves reasonable precision and can thus be used to complement existing approaches to filter for high confidence circRNAs and obtain their full length sequences. ACValidator is a novel bioinformatics workflow that can be used to validate candidate circRNAs of interest in silico and help distinguish true positive candidates. When different algorithms identify disparate circRNA candidates, ACValidator will be helpful in narrowing down specific circRNAs of interest to identify those associated with higher confidence. Further, this can also serve as a circRNA candidate selection tool for wet lab validations. ACValidator performs more optimally when a higher number of supporting reads are available and can be used for both RNAse R treated and non-treated samples. These two

novel assembly based approaches bring a new perspective to circRNA detection methodologies.

Finally, we applied these two approaches, along with existing tools, to systematically characterize circRNA expression in five functionally distinct regions of healthy aged human brain – the BC, IP, MG, OC and SF. We also identified DE circRNAs in each region and their predicted impact on transcriptional networks. Genes from these networks were significantly enriched in nervous system developmental functions and pathways such as brain size and formation, development, morphology and proliferation of neurons, synaptic plasticity, proliferation of neuroglia and long-term potentiation. High confidence circRNAs, detected by both DeFuCir and existing approaches, map to *RIMS1* in the BC, OC and SF, *RIMS2* in the MG and *MLIP* in the IP. Further, we were able to in silico validate approximately 45% of DE circRNAs identified by DESeq2 and more than 80% of circRNAs called by all six tools in all samples.

2. Limitations

Although we were able to perform the aforementioned analyses, we are limited by a few caveats. In Chapter 2, the cell-specific transcriptomic data analyzed was not treated with RNAse R, and hence there is a larger pool of transcripts, majority of which are linear. As a result, we may not have efficiently captured astrocytic circRNAs in our samples. On the other hand, in Chapter 4, we use RNase R treatment to enrich for circRNAs in our samples. Although this treatment is essential in circRNA detection, it is not well understood if this depletion step introduces any biases in circRNA detection and

whether this step has the potential to deplete circRNAs. As we gain further insights into circRNA biology and their sensitivity/resistance to RNAse R, we may be able to tackle the problem of circRNA enrichment more efficiently. In Chapter 3, we were primarily limited by the lack of a gold standard circRNA reference dataset for evaluation of our approaches. As a result, we rely on simulation datasets, which are further based on informatically predicted circRNAs detected by various studies and deposited in circBase. Moreover, the small list of experimentally validated circRNAs from various studies (N = 282) is derived from different cell and tissue types and hence may not be directly comparable to our datasets. In order to implement an unbiased, assembly based circRNA detection methodology, DeFuCir uses minimal filtering thresholds, and hence, repeat regions or single nucleotide variants (SNVs) in the region overlapping between the soft-clipped portion and the lowest segment of a contig are currently not accounted for in our pipeline. These repeat regions may represent potential false positives and hence will be handled in the next version of the DeFuCir workflow using the RepeatMasker program [101]. This program and identify and mask repeated regions from analysis. Further, since our assembly analysis relies on overlapping read's coverage over the regions of interest, we are limited in our ability to capture as well as *in silico* validate circRNAs whose abundance might be below detection levels in samples, particularly when they are not enriched for circRNAs.

3. Future Directions

Whole transcriptome analysis is necessary for understanding how altered gene expression contributes to complex diseases such as AD and cancer. Recent revelations on a widespread number of coding and non-coding species of RNA enable a more comprehensive system level perspective of transcriptional regulation. Such understanding is needed to decipher the molecular underpinnings of basic cellular mechanisms as well as developing improved therapeutic strategies for rare and debilitating diseases. Moving toward this important goal, the next step of our research is to carry out whole transcriptome analysis of the brain region samples described in Chapter 4 by sequencing small RNAs, mRNAs, and long non-coding RNAs. Such analysis will both help characterize other RNA species as well as enable us to better evaluate the significance of circRNAs in the context of the whole transcriptome and its regulatory networks. Specifically, quantifying and comparing their expression will be performed to evaluate the abundance of each species relative to the whole transcriptome. Further, we will continue to develop DeFuCir and ACValidator as needed to account for new insights into circRNA biology. We will also include a module for handling repeat regions in the genome for DeFuCir in order to eliminate potential false positives.

Given our identification of circRNAs that are unique and common across the brain regions we investigated, further functional studies on these candidates will help lend insights into their biological significance and clarify whether they truly represent an additional layer of transcriptional regulation in the brain. Additionally, comparing data from the current study with data from healthy younger individuals as well as affected

elderly individuals will help us to understand whether they are specific to age or other conditions. Lastly, the workflows developed in this thesis are applicable to any non-polyA selected RNAseq dataset and can thus be utilized to characterize circRNA expression across various sample types and diseases.

REFERENCES

1. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO: **Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types**. *PloS one* 2012, **7**(2):e30733.

2. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE: **Circular RNAs are abundant, conserved, and associated with ALU repeats**. *RNA (New York, NY)* 2013, **19**(2):141-157.

3. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M: **Circular RNAs are a large class of animal RNAs with regulatory potency**. *Nature* 2013, **495**(7441):333-338.

4. Sanger HL, Klotz G, Riesner D, Gross HJ, Kleinschmidt AK: **Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures**. *Proceedings of the National Academy of Sciences of the United States of America* 1976, **73**(11):3852-3856.

5. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, Kinzler KW, Vogelstein B: **Scrambled exons**. *Cell* 1991, **64**(3):607-613.

6. Capel B, Swain A, Nicolis S, Hacker A, Walter M, Koopman P, Goodfellow P, Lovell-Badge R: **Circular transcripts of the testis-determining gene Sry in adult mouse testis**. *Cell* 1993, **73**(5):1019-1030.

7. Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO: **Cell-type specific features of circular RNA expression**. *PLoS genetics* 2013, **9**(9):e1003777.

8. Chen C-y, Sarnow P: **Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs**. *Science* 1995, **268**(5209):415.

9. Perriman R, Ares M, Jr.: **Circular mRNA can direct translation of extremely long repeating-sequence proteins in vivo**. *RNA (New York, NY)* 1998, **4**(9):1047-1054.

10. Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, Roslan S, Schreiber AW, Gregory PA, Goodall GJ: **The RNA binding protein quaking regulates formation of circRNAs**. *Cell* 2015, **160**(6):1125-1134.

11. Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C: **Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals**. *Cell reports* 2015, **10**(2):170-177.

12. Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L: **Exon-intron circular RNAs regulate transcription in the nucleus**. *Nature structural & molecular biology* 2015, **22**(3):256-264.

13. Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, Zhu S, Yang L, Chen L-L: **Circular intronic long noncoding RNAs**. *Molecular cell* 2013, **51**(6):792-806.

14. Lu Z, Filonov GS, Noto JJ, Schmidt CA, Hatkevich TL, Wen Y, Jaffrey SR, Matera AG: **Metazoan tRNA introns generate stable circular RNAs in vivo**. *RNA* 2015, **21**(9):1554-1565.

15. Hansen TB, Wiklund ED, Bramsen JB, Villadsen SB, Statham AL, Clark SJ, Kjems J: **miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA**. *The EMBO journal* 2011, **30**(21):4414-4422.

16. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J: **Natural RNA circles function as efficient microRNA sponges**. *Nature* 2013, **495**(7441):384-388.

17. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S: **circRNA biogenesis competes with pre-mRNA splicing**. *Molecular cell* 2014, **56**(1):55-66.

18. Du WW, Yang W, Chen Y, Wu Z-K, Foster FS, Yang Z, Li X, Yang BB: **Foxo3 circular RNA promotes cardiac senescence by modulating multiple factors associated with stress and senescence responses**. *European heart journal* 2016, **38**(18):1402-1412.

19. Du WW, Yang W, Liu E, Yang Z, Dhaliwal P, Yang BB: **Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2**. *Nucleic acids research* 2016, **44**(6):2846-2858.

20. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, Hanan M, Behm M, Bartok O, Ashwal-Fluss R: **Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed**. *Molecular cell* 2015, **58**(5):870-885.

21. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.

22. You X, Vlatkovic I, Babic A, Will T, Epstein I, Tushev G, Akbalik G, Wang M, Glock C, Quedenau C: **Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity**. *Nature neuroscience* 2015.

23. Memczak S, Papavasileiou P, Peters O, Rajewsky N: **Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood**. *PLoS ONE* 2015, **10**(10):e0141214.

24. Tan WL, Lim BT, Anene-Nzelu CG, Ackers-Johnson M, Dashi A, See K, Tiang Z, Lee DP, Chua W, Luu TD: **A landscape of circular RNA expression in the human heart**. *Cardiovascular Research* 2016:cvw250.

25. Maass PG, Glažar P, Memczak S, Dittmar G, Hollfinger I, Schreyer L, Sauer AV, Toka O, Aiuti A, Luft FC: **A map of human circular RNAs in clinically relevant tissues**. *Journal of Molecular Medicine* 2017, **95**(11):1179-1189.

26. Xu T, Wu J, Han P, Zhao Z, Song X: **Circular RNA expression profiles and features in human tissues: a study using RNA-seq data**. *BMC Genomics* 2017, **18**(Suppl 6):680.

27. Guo JU, Agarwal V, Guo H, Bartel DP: **Expanded identification and characterization of mammalian circular RNAs**. *Genome biology* 2014, **15**(7):1.

28. Gruner H, Cortés-López M, Cooper DA, Bauer M, Miura P: **CircRNA accumulation in the aging mouse brain**. *Scientific reports* 2016, **6**:38907.

29. Werfel S, Nothjunge S, Schwarzmayr T, Strom T-M, Meitinger T, Engelhardt S: **Characterization of circular RNAs in human, mouse and rat hearts**. *Journal of Molecular and Cellular Cardiology* 2016, **98**:103-107.

30. Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai EC: **Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation**. *Cell reports* 2014, **9**(5):1966-1980.

31. Cortés-López M, Gruner MR, Cooper DA, Gruner HN, Voda A-I, van der Linden AM, Miura P: **Global accumulation of circRNAs during aging in Caenorhabditis elegans**. *BMC Genomics* 2018, **19**(1):8.

32. Lukiw W: **Circular RNA (circRNA) in Alzheimer's disease (AD)**. *Frontiers in genetics* 2013, **4**:307.

33. Zhu J, Ye J, Zhang L, Xia L, Hu H, Jiang H, Wan Z, Sheng F, Ma Y, Li W: **Differential expression of circular RNAs in glioblastoma multiforme and its correlation with prognosis**. *Translational oncology* 2017, **10**(2):271-279.

34. Bachmayr-Heyda A, Reiner AT, Auer K, Sukhbaatar N, Aust S, Bachleitner-Hofmann T, Mesteri I, Grunt TW, Zeillinger R, Pils D: **Correlation of circular RNA abundance with proliferation-exemplified with colorectal and ovarian**

cancer, idiopathic lung fibrosis, and normal human tissues. *Scientific reports* 2015, **5**.

35. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE: **Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk**. *PLoS Genet* 2010, **6**(12):e1001233.

36. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM *et al*: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery**. *Nucleic acids research* 2010, **38**(18):e178.

37. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357.

38. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. *arXiv preprint arXiv:13033997* 2013.

39. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome biology* 2009, **10**(3):R25.

40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

41. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome biology* 2013, **14**(4):R36.

42. Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, Yang L: **Complementary sequence-mediated exon circularization**. *Cell* 2014, **159**(1):134-147.

43. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts**. *Genome biology* 2011, **12**(8):R72.

44. Cheng J, Metge F, Dieterich C: **Specific identification and quantification of circular RNAs from sequencing data**. *Bioinformatics* 2016, **32**(7):1094-1096.

45. Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, Jiang C, Parast MM, Murry CE, Laurent LC, Salzman J: **Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development**. *Genome biology* 2015, **16**:126-015-0690-0695.

46. Gao Y, Wang J, Zhao F: **CIRI: an efficient and unbiased algorithm for de novo circular RNA identification**. *Genome biology* 2015, **16**:4-014-0571-0573.

47. Hansen TB, Venø MT, Damgaard CK, Kjems J: **Comparison of circular RNA prediction tools**. *Nucleic Acids Res* 2016, **44**(6):e58.

48. Zeng X, Lin W, Guo M, Zou Q: **A comprehensive overview and evaluation of circular RNA detection tools**. *PLoS computational biology* 2017, **13**(6):e1005420.

49. Szabo L, Salzman J: **Detecting circular RNAs: bioinformatic and experimental challenges**. *Nature reviews Genetics* 2016, **17**(11):679-692.

50. Glazar P, Papavasileiou P, Rajewsky N: **circBase: a database for circular RNAs**. *RNA (New York, NY)* 2014, **20**(11):1666-1670.

51. Ghosal S, Das S, Sen R, Basak P, Chakrabarti J: **Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits**. *Frontiers in genetics* 2013, **4**.

52. Liu Y-C, Li J-R, Sun C-H, Andrews E, Chao R-F, Lin F-M, Weng S-L, Hsu S-D, Huang C-C, Cheng C: **CircNet: a database of circular RNAs derived from transcriptome sequencing data**. *Nucleic acids research* 2015:gkv940.

53. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H: **starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data**. *Nucleic Acids Research* 2014, **42**(Database issue):D92-D97.

54. Xia S, Feng J, Chen K, Ma Y, Gong J, Cai F, Jin Y, Gao Y, Xia L, Chang H *et al*: **CSCD: a database for cancer-specific circular RNAs**. *Nucleic Acids Research* 2018, **46**(D1):D925-D929.

55. Xia S, Feng J, Lei L, Hu J, Xia L, Wang J, Xiang Y, Liu L, Zhong S, Han L *et al*: **Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes**. *Briefings in Bioinformatics* 2017, **18**(6):984-992.

56. Chen X, Han P, Zhou T, Guo X, Song X, Li Y: **circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations**. *Sci Rep* 2016, **6**:34985.

57. Dudekula DB, Panda AC, Grammatikakis I, De S, Abdelmohsen K, Gorospe M: **CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs**. *RNA biology* 2016, **13**(1):34-42.

58. Li F, Zhang L, Li W, Deng J, Zheng J, An M, Lu J, Zhou Y: **Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/beta-catenin pathway**. *Oncotarget* 2015, **6**(8):6001-6013.

59. Li P, Chen S, Chen H, Mo X, Li T, Shao Y, Xiao B, Guo J: **Using circular RNA as a novel type of biomarker in the screening of gastric cancer**. *Clinica Chimica Acta* 2015, **444**:132-136.

60. Parpura V, Verkhratsky A: **Homeostatic function of astrocytes: Ca(2+) and Na(+) signalling**. *Translational neuroscience* 2012, **3**(4):334-344.

61. Jensen CJ, Massie A, De Keyser J: **Immune players in the CNS: the astrocyte**. *Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology* 2013, **8**(4):824-839.

62. Tsacopoulos M, Magistretti PJ: **Metabolic coupling between glia and neurons**. *Journal of Neuroscience* 1996, **16**(3):877-885.

63. Pellerin L, Magistretti PJ: **Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization**. *Proceedings of the National Academy of Sciences* 1994, **91**(22):10625-10629.

64. Sekar S, McDonald J, Cuyugan L, Aldrich J, Kurdoglu A, Adkins J, Serrano G, Beach TG, Craig DW, Valla J *et al*: **Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes**. *Neurobiol Aging* 2015, **36**(2):583-591.

65. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in Drosophila**. *Genome biology* 2003, **5**(1):R1.

66. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes**. *Rna* 2004, **10**(10):1507-1517.

67. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic Acids Research* 2014, **42**(D1):D68-D73.

68. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

69. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ *et al*: **miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database**. *Nucleic Acids Res* 2016, **44**(D1):D239-247.

70. Agarwal V, Bell GW, Nam J-W, Bartel DP: **Predicting effective microRNA target sites in mammalian mRNAs**. *elife* 2015, **4**.

71. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.

72. Farina C, Aloisi F, Meinl E: **Astrocytes are active players in cerebral innate immunity**. *Trends in immunology* 2007, **28**(3):138-145.

73. Gadani SP, Cronk JC, Norris GT, Kipnis J: **IL-4 in the Brain: A Cytokine To Remember**. *The Journal of Immunology* 2012, **189**(9):4213-4219.

74. Hansen TB, Kjems J, Damgaard CK: **Circular RNA and miR-7 in cancer**. *Cancer Res* 2013, **73**(18):5609-5612.

75. Dropcho EJ, Chen YT, Posner JB, Old LJ: **Cloning of a brain protein identified by autoantibodies from a patient with paraneoplastic cerebellar degeneration**. *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84**(13):4552-4556.

76. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, Caselli RJ, Kukull WA, McKeel D, Morris JC *et al*: **Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain**. *Physiol Genomics* 2007, **28**(3):311-322.

77. Chen GH, Wang YJ, Qin S, Yang QG, Zhou JN, Liu RY: **Age-related spatial cognitive impairment is correlated with increase of synaptotagmin 1 in dorsal hippocampus in SAMP8 mice**. *Neurobiol Aging* 2007, **28**(4):611-618.

78. Zhou X, Sun L, Bracko O, Choi JW, Jia Y, Nana AL, Brady OA, Hernandez JCC, Nishimura N, Seeley WW *et al*: **Impaired prosaposin lysosomal trafficking in frontotemporal lobar degeneration due to progranulin mutations**. *Nature communications* 2017, **8**:15277.

79. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R: **Systematic evaluation of spliced alignment programs for RNA-seq data**. *Nature methods* 2013, **10**(12):1185-1191.

80. **FastQC: A Quality Control tool for High Throughput Sequence Data** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]

81. **Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences** [https://github.com/lh3/seqtk]

82. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome research* 2002, **12**(6):996-1006.

83. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics (Oxford, England)* 2010, **26**(6):841-842.

84. Beach TG, Adler CH, Sue LI, Serrano G, Shill HA, Walker DG, Lue L, Roher AE, Dugger BN, Maarouf C: **Arizona study of aging and neurodegenerative disorders and brain and body donation program**. *Neuropathology* 2015, **35**(4):354-389.

85. Newman AM, Bratman SV, Stehr H, Lee LJ, Liu CL, Diehn M, Alizadeh AA: **FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution**. *Bioinformatics* 2014, **30**(23):3390-3393.

86. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L: **CREST maps somatic structural variation in cancer genomes with base-pair resolution**. *Nature methods* 2011, **8**(8):652.

87. Yang R, Nelson AC, Henzler C, Thyagarajan B, Silverstein KA: **ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly**. *Genome medicine* 2015, **7**(1):127.

88. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nature protocols* 2013, **8**(8):1494.

89. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database C: **The Sequence Read Archive**. *Nucleic Acids Research* 2011, **39**(Database issue):D19-D21.

90. Guarnerio J, Bezzi M, Jeong Jong C, Paffenholz Stella V, Berry K, Naldini Matteo M, Lo-Coco F, Tay Y, Beck Andrew H, Pandolfi Pier P: **Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations**. *Cell*, **165**(2):289-302.

91. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K *et al*: **Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics**. *Molecular & cellular proteomics : MCP* 2014, **13**(2):397-406.

92. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET *et al*: **A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project**. *Biopreservation and Biobanking* 2015, **13**(5):311-319.

93. Schmidt-Ullrich R, Memet S, Lilienbaum A, Feuillard J, Raphael M, Israel A: **NF-kappaB activity in transgenic mice: developmental regulation and tissue specificity**. *Development (Cambridge, England)* 1996, **122**(7):2117-2128.

94. Koulich E, Nguyen T, Johnson K, Giardina CA, D'mello SR: **NF‐κB is involved in the survival of cerebellar granule neurons: association of IκB phosphorylation with cell survival**. *Journal of neurochemistry* 2001, **76**(4):1188-1198.

95. O'Neill LA, Kaltschmidt C: **NF-kB: a crucial transcription factor for glial and neuronal cell function**. *Trends in neurosciences* 1997, **20**(6):252-258.

96. Mattson MP, Culmsee C, Yu Z, Camandola S: **Roles of nuclear factor κB in neuronal survival and plasticity**. *Journal of neurochemistry* 2000, **74**(2):443-456.

97. Li W, Qin W, Liu H, Fan L, Wang J, Jiang T, Yu C: **Subregions of the human superior frontal gyrus and their connections**. *NeuroImage* 2013, **78**:46-58.

98. De A: **Wnt/Ca2+ signaling pathway: a brief overview**. *Acta Biochimica et Biophysica Sinica* 2011, **43**(10):745-756.

99. Liang WS, Reiman EM, Valla J, Dunckley T, Beach TG, Grover A, Niedzielko TL, Schneider LE, Mastroeni D, Caselli R: **Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons**. *Proceedings of the National Academy of Sciences* 2008, **105**(11):4441-4446.

100. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Ramsey K, Caselli RJ, Kukull WA, McKeel D, Morris JC *et al*: **Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set**. *Physiological Genomics* 2008, **33**(2):240-256.

101. Smit A, Hubley R, Green P: **2013–2015. RepeatMasker Open-4.0**. In.; 2013.

APPENDIX A

PERMISSION TO USE PUBLISHED MATERIAL

Content in Chapter 2 is published on the pre-print server Biorxiv, and is also accepted for publication in the peer reviewed journal *BMC Genomics*. All co-authors have granted permission to use the publication content in Chapter 2 of this dissertation.