

**A NOVEL BIOINFORMATICS APPROACH TO CHARACTERIZE AND  
INTEGRATE MESSENGER RNAs, CIRCULAR RNAs AND MICRO RNAs**

A THESIS  
SUBMITTED TO THE FACULTY OF  
UNIVERSITY OF MINNESOTA  
BY

ASHA A. NAIR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

DR. KRISHNA R. KALARI and DR. SUBBAYA SUBRAMANIAN, ADVISORS

APRIL, 2018

© ASHA NAIR 2018

## Acknowledgements

First and foremost, I would like to acknowledge my phenomenal advisors – Dr. Krishna R. Kalari and Dr. Subbaya Subramanian, for providing me with the opportunity to work on this wonderful project. I am deeply thankful to them for their faith and patience in me throughout this journey and appreciate the valuable time that they contributed to making my Ph.D. experience so inspiring and productive. They showed me how good science is done. I thank Dr. Kalari for setting the excellent example of a successful woman researcher and for her encouragement, guidance and precious support throughout my research. I also thank Dr. Subramanian for being so enlightening and motivational and always accommodating our long-distance Skype meetings into his busy schedule. It has been an honor to be your Ph.D. student.

I am grateful to my committee members, Dr. Chad Myers and Dr. Jasmine Foo for their brilliant insights and guidance. I am thankful to them for their valuable feedback on my proposal writing, dissertation and project work. Thank you for having me as your student; your support has been invaluable, and I very much appreciate that.

I gratefully acknowledge Dr. Claudia Neuhauser for offering me her academic and moral support right from when I started. Dr. Neuhauser's role in my career

has been instrumental, and I will always treasure her encouraging words during the tough times of my Ph.D. pursuit.

Last but never the least; I would like to take this opportunity to thank all the people who helped me along the way, my colleagues, friends, and family, all of whom were invaluable. I wish to specially thank Dr. Xiaojia Tang, Dr. Kevin Thompson and Dr. Jaime Davila for being such great colleagues and supporting me whenever I needed it. Being a part-time Ph.D. student and a full-time employee has been challenging, but working at an amazing organization like Mayo Clinic has made this feat possible.

## **Dedication**

This dissertation is dedicated to our loving parents who encouraged and believed in me throughout this journey and my dear husband whose timeless support and love has made this accomplishment possible.

## **Abstract**

High-throughput Next Generation RNA sequencing (RNA-Seq) technology is affluent with information about the transcriptome, which includes both protein-coding and multiple non-coding regions. In a diseased state, complex interactions between these regions can go awry. Identification of such interactions is critical to translate the underlying biology of the transcriptome, especially for lethal diseases such as cancer. The field of bioinformatics is currently deficient in workflows that can analyze both coding and non-coding regions together efficiently, to identify disease-specific interactions.

In this dissertation, I developed three coherent bioinformatics solutions that aim to address these shortcomings in RNA-Seq. First, a comprehensive workflow called MAPR-Seq was developed to analyze and report various features of protein-coding messenger RNAs. Second, a workflow for non-coding circular RNAs, called Circ-Seq, was developed to identify, quantify and annotate expressed circular RNAs. Third, an integration workflow called ReMlx was developed to identify microRNA response elements (MREs) and integrate them with the different types of RNAs (messenger RNAs, circular RNAs, and microRNAs).

Collectively, the three workflows were applied to the largest cohort of breast cancer samples (n=885) from The Cancer Genome Atlas (TCGA). Based on the results obtained from these workflows, I present several key findings that are pertinent to breast cancer. I show that circular RNAs may be a marker for tumor proliferation in estrogen response positive (ER+) breast cancer subtype. I also show how triple negative (TN) breast cancer subtype-specific MRE signatures of messenger RNA – microRNA interactions can be obtained using RNA-Seq data, which has not been explored to date and thus, is a novel undertaking. In the end, my results highlight candidate messenger RNAs, circular RNAs and microRNAs that are found to be associated with MAPK and PI3K/AKT signaling cascades in TN breast cancer subtype.

In general, the developed bioinformatics solutions can also be applied to RNA-Seq data of other cancer subtypes and diseases to identify unique messenger RNA – microRNA – circular RNA candidates that could be promising diagnostic targets towards improving treatment options for complex diseases.

# Table of Contents

<b>Acknowledgements</b> .....	i
<b>Dedication</b> .....	iii
<b>Abstract</b> .....	iv
<b>Table of Contents</b> .....	vi
<b>List of Tables</b> .....	viii
<b>List of Figures</b> .....	ix
<b>Chapter 1: Introduction</b> .....	1
1.1 Evolutionary expansion of the genome.....	1
1.2 Current landscape of the human transcriptome.....	2
1.3 Types of RNA considered in this dissertation .....	4
1.3.1 Messenger RNAs .....	4
1.3.2 MicroRNAs .....	5
1.3.3 Circular RNAs.....	5
1.4 Cross-talk between different RNA types .....	6
1.5 Importance of studying RNA interactions in Breast Cancer .....	8
1.6 Availability of high-throughput NGS datasets .....	9
1.7 Challenges and Motivation .....	11
1.8 Outline of the chapters .....	13
<b>Chapter 2: MAP-RSeq – Mayo Analysis Pipeline for RNA Sequencing</b> .....	16
2.1 Background .....	16
2.2 Availability and requirements.....	17
2.3 Implementation.....	18
2.3.1 Virtual machine.....	18
2.3.2 Sun grid engine .....	20
2.4 Results .....	21
2.4.1 Gene expression and exon expression read counts .....	25
2.4.2 Differential expression .....	29
2.4.3 Expressed SNVs (eSNVs) from RNA-Seq.....	29
2.4.4 Fusion transcript detection.....	32
2.4.5 Summarization of data and final report .....	34
2.5 Conclusions.....	34
<b>Chapter 3: CircRNAs and their associations with breast cancer subtypes</b> . 36	
3.1 Introduction .....	36
3.2 Results .....	39



3.2.1	Circ-Seq: an automated workflow for circRNA identification .....	39
3.2.2	Identification of circRNAs in breast cancer cell lines.....	40
3.2.3	Validation of circRNA in MCF7 breast cancer cells.....	41
3.2.4	Presence of circRNAs in TCGA breast cancer transcriptomes .....	43
3.2.4.1	Breast cancer subtype analysis.....	44
3.2.4.2	Paired normal-adjacent tissue analysis .....	50
3.3	Methods .....	55
3.3.1	Circ-Seq workflow.....	55
3.3.2	TCGA breast cancer transcriptome data .....	58
3.3.3	TCGA breast tumor and normal-adjacent samples and normal breast mammary tissue from GTEx .....	58
3.3.4	Breast cancer cell lines.....	59
3.3.5	Pathway analysis for tumor-specific circRNAs.....	59
3.3.6	CircRNA validation .....	60
3.4	Discussion.....	61
<b>Chapter 4: ReMlx - A novel bioinformatics approach to integrate mRNA- microRNA interactions using MRE frequencies from RNA-Seq data .....</b>		<b>66</b>
4.1	Introduction.....	66
4.2	Results .....	69
4.2.1	Application of ReMlx to triple-negative tumor and normal-adjacent pairs....	69
4.2.2	614 MREs associated to 272 genes and 198 microRNAs.....	71
4.2.3	MAPK and PI3K-AKT signaling identified among top pathways .....	74
4.2.5	CircRNAs associated with MAPK and PI3K-AKT pathways.....	78
4.2.6	mRNA-microRNA-circRNA interacting candidates in MAPK and PI3K-AKT pathways for TN breast cancer .....	80
4.3	Methods.....	83
4.3.1	ReMlx – a novel methodology to compute MRE frequency from RNA-Seq data.....	83
4.3.2	MRE frequency analysis from RNA-Seq data .....	84
4.3.3	3'UTR definitions obtained from TargetScan.....	87
4.3.4	RNA-Seq and microRNA-Seq data from TCGA .....	87
4.3.5	Statistical analyses on MRE sites and activated pathway identification .....	88
4.3.6	Pathway analysis for canonical pathways .....	88
4.4	Discussion .....	89
<b>Chapter 5: Conclusions and Discussion .....</b>		<b>93</b>
<b>Bibliography .....</b>		<b>101</b>
<b>Appendix .....</b>		<b>111</b>
6.1	Permissions .....	111

## List of Tables

<b>Table 1:</b> MAP-RSeq installation and runtime for QuickStart virtual machine .....	19
<b>Table 2:</b> MAP-RSeq installation and runtime in a Linux environment .....	20
<b>Table 3:</b> Wall clock times to run MAP-RSeq at different read counts .....	21
<b>Table 4:</b> Alignment statistics from MAP-RSeq using a simulated dataset from BEERS .....	27
<b>Table 5:</b> Number of circRNAs identified in breast cell lines using the Circ-Seq workflow .....	41
<b>Table 6:</b> Summary of breast tumors, adjacent tissues, and tumor-specific circRNAs in sequence data made available by the cancer genome atlas .....	45
<b>Table 7:</b> Gene-microRNA pairs with distinct TN-specific MRE sites that are part of the MAPK and PI3K-AKT pathways.....	78
<b>Table 8:</b> Intra-gene circRNAs identified in MAPK and PI3K-AKT pathways.. .....	79

## List of Figures

<b>Figure 1:</b> Evolutionary expansion of non-coding regions in various organisms .....	2
<b>Figure 2:</b> Endogenous competition between mRNAs and circRNAs for a common pool of microRNAs.....	7
<b>Figure 3:</b> Flowchart of the MAP-RSeq workflow. High-level representation of the MAP-RSeq workflow for processing RNA-Seq data.....	23
<b>Figure 4:</b> Screenshot output report (html) of MAP-RSeq. An example screenshot report of the MAP-RSeq output file. ....	24
<b>Figure 5:</b> Correlation of gene counts reported by MAP-RSeq in comparison to counts simulated by BEERS.....	26
<b>Figure 6:</b> Screenshots of gene and exon expression reports by MAP-RSeq.....	28
<b>Figure 7:</b> Screenshot of a MAP-RSeq VCF file after VQSR annotation.....	30
<b>Figure 8:</b> Examples of SNVs called in RNA and DNA data for NA07347.. ....	31
<b>Figure 9:</b> Fusion transcripts reported by MAP-RSeq. An example of the fusion transcripts output file from MAP-RSeq workflow. ....	33
<b>Figure 10:</b> Validation of a circRNA at locus chr14:102,466,325–102,500,789. (A) circRNA was amplified by divergent primers using total RNA but not genomic DNA (gDNA). GAPDH was used as a control. (B) Head-to-tail splicing was confirmed by Sanger Sequencing. ....	43
<b>Figure 11:</b> TCGA tumor-specific circRNAs also found in breast cell lines. (A) overlap of circRNAs between different subtypes for breast cell lines, (B) overlap of TN and ER+ tumor-specific circRNAs between TCGA and cell lines.....	47
<b>Figure 12:</b> Tumor-specific circRNAs common and unique to TN, ER+ and HER2+ subtypes and the top canonical pathways associated with each subtype. ....	49
<b>Figure 13:</b> Increased number of circRNAs in normal breast samples compared to breast tumor subtypes in TCGA.....	51
<b>Figure 14:</b> Lower number of circRNAs as gene proliferation increases in ER+ tumor samples .....	53

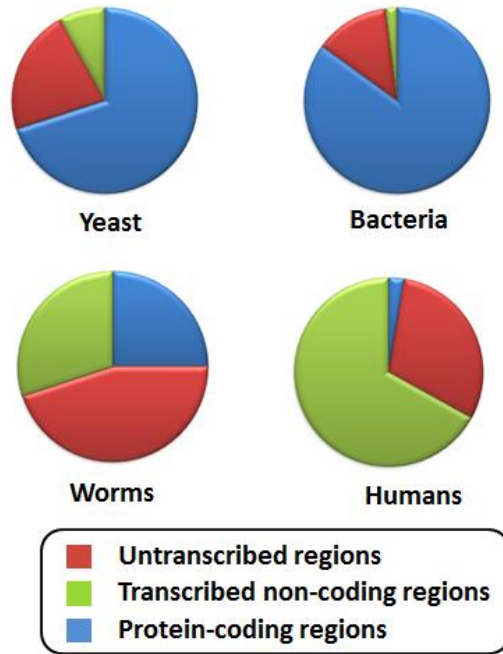
<b>Figure 15:</b> (A) Luminal A and Luminal B tumor samples show distinct separation based on their circRNA numbers when plotted against tumor proliferation, (B) Unsupervised hierarchical clustering analysis shows separation of Luminal A and Luminal B tumor and adjacent samples based on their circRNA numbers.....	55
<b>Figure 16:</b> Circ-Seq bioinformatics workflow flowchart.....	57
<b>Figure 17:</b> Heatmap of 614 TN-tumor specific MREs.....	71
<b>Figure 18:</b> Un-supervised clustering and heatmap representation 614 TN-tumor specific MREs and their associated genes and microRNAs..	72
<b>Figure 19:</b> Heatmap of gene expression for 272 genes..	73
<b>Figure 20:</b> MAPK signaling pathway..	75
<b>Figure 21:</b> PI3K-AKT signaling pathway...	76
<b>Figure 22:</b> MAPK endogenous RNA network.....	81
<b>Figure 23:</b> PI3K-AKT endogenous RNA network..	82
<b>Figure 24:</b> Hypothetical representation of MRE frequency counting using RNA-Seq data..	84
<b>Figure 25:</b> Flowchart representation of MRE frequency quantification from RNA-Seq BAM.....	85
<b>Figure 26:</b> Permission from BMC Bioinformatics to reproduce MAP-RSeq publication .....	112
<b>Figure 27:</b> Snapshot of publication for MAP-RSeq workflow.....	113
<b>Figure 28:</b> Permission from Oncotarget to reproduce circular RNA publication ....	114
<b>Figure 29:</b> Creative Commons license agreement used by Oncotarget .....	115
<b>Figure 30:</b> Snapshot of publication for Circ-Seq workflow and circular RNA associations found in breast cancer.....	116

# Chapter 1: Introduction

## 1.1 Evolutionary expansion of the genome

Historically, scientists and researchers have been keen on studying genomic size, evolution and functional significance in numerous species [1-3]. During the era of the 1950s, it was assumed that genomic size correlates with organism complexity [4]. Scientists strongly believed that humans are one of the most complex species on earth and thus would have the largest genome with a maximum number of genes. However, this theory was proven wrong when it was discovered in 1971 that lower animals such as salamander have a genome 15 times larger than that of humans [5]. This was a troubling paradox for many years. Later, experimental advances revealed that the genome is comprised of protein-coding and non-coding regions [6]. It was found that while a small subset is protein-coding, the non-coding content in the genomes of these different species varied by several folds [7, 8]. Since 2003, with the revolution of sequencing technologies and formation of the Encyclopedia of DNA Elements (ENCODE) consortium led by the US National Human Genome Research Institute (NHGRI), the mystery began to unfold by itself. It became surprisingly evident that non-coding regions of the genome also undergo transcription and, as shown in Figure 1, the number of transcribed non-coding regions increased for higher species [9, 10]. Thus it was clear, that the non-coding content has

undergone evolutionary expansion, suggesting their potential to correlate with organismal complexity.



**Figure 1:** Evolutionary expansion of non-coding regions in various organisms

## 1.2 Current landscape of the human transcriptome

As mentioned above, one of the many surprises that came from the ENCODE project was that over 70% of the non-coding regions in the human genome are transcribed and that only 2% of the genome encodes for proteins [10]. In 2012, the GENCODE project (Encyclopedia of genes and gene variants) emerged, leading to the categorization of transcribed, non-coding RNA molecules into several classes based on their characteristics and functional implications in the

genome [11]. It is worth a mention that even at present, the research community continues to discover new, unfound members of the non-coding RNA family as their presence continues to become evident using the latest and greatest sequencing technologies coupled with bioinformatics techniques.

Non-coding RNAs have been broadly classified into categories such as long non-coding RNAs (lncRNAs), small non-coding RNAs (sncRNAs), pseudogenes and an unclassified category of unprocessed transcripts (which consist of less understood genes that eventually get characterized based on on-going research and validation techniques) – for example, circular RNAs. The lncRNAs are defined as RNA molecules longer than 200 bases which exert pre- and post-transcriptional regulatory effects on their messenger RNA (mRNA) counterparts [12]. The sncRNAs are typically 25 – 30 bases long which have a variety of functional and regulatory mechanisms and thus have been distinguished further into many more specific types such as transfer RNAs, ribosomal RNAs, small nucleolar RNAs, microRNAs, small interfering RNAs, etc. [13, 14]. Pseudogenes are non-functional relatives of mRNAs that have lost their protein-coding capacity due to various alterations [15]. Finally, circRNAs are non-canonical back-spliced by-products of transcription that form the most recent addition (discovered in 2013 in humans) to the family of noncoding RNAs [16].

In summary, according to the latest GENCODE version (v19) of human genome build hg19, the human transcriptome consists of 58,050 genes which comprise of 35% mRNAs, 24% lncRNAs, 15% sncRNAs, 24% pseudogenes and 2% unclassified/unprocessed transcripts (which includes circRNAs).

### **1.3 Types of RNA considered in this dissertation**

In this dissertation, I will focus on protein-coding mRNAs and two types of non-coding RNAs, namely, microRNAs and circRNAs. This dissertation is an integrated study of these three RNA types.

#### **1.3.1 Messenger RNAs**

Messenger RNAs (mRNAs) are a well-studied class of RNA molecules that carry genetic codes from DNA in the nucleus to sites for protein synthesis in the cytoplasm. Structurally, mRNAs comprises of the following in the given order: 5' cap, 5' untranslated region (UTR), coding exons, 3' untranslated region, polyadenylated (poly-A) tail. The poly-A tail protects the mRNA from degradation by exonucleases. However, the untranslated regions, especially the 3'UTR, can serve as hot spots for binding of specific non-coding RNAs called microRNAs (explained below) to mediate degradation of the expressed gene.



### **1.3.2 MicroRNAs**

MicroRNAs, a distinct category of sncRNAs, are highly conserved, single-stranded RNA molecules of approximately 22 bases in length. Mature microRNAs are formed based on a unique transcription process. Export of microRNAs from the nucleus to the cytoplasm is performed by employing the Drosha, Exportin-5 and Dicer enzymes. Once mature, microRNAs accomplish their regulatory functions through the RNA-induced silencing complex (RISC) [17]. MicroRNAs activate and guide the RISC complex towards their target mRNAs to regulate gene expression. The RISC complex recognizes microRNA response elements (MRE) present on the 3'UTR of target mRNAs for complementary base-pairing with microRNAs. The degree and nature of the complementarity between the microRNA and its target gene determines the gene silencing mechanism, i.e., whether the gene undergoes mRNA degradation (imperfect pairing) or translation inhibition (perfect pairing).

### **1.3.3 Circular RNAs**

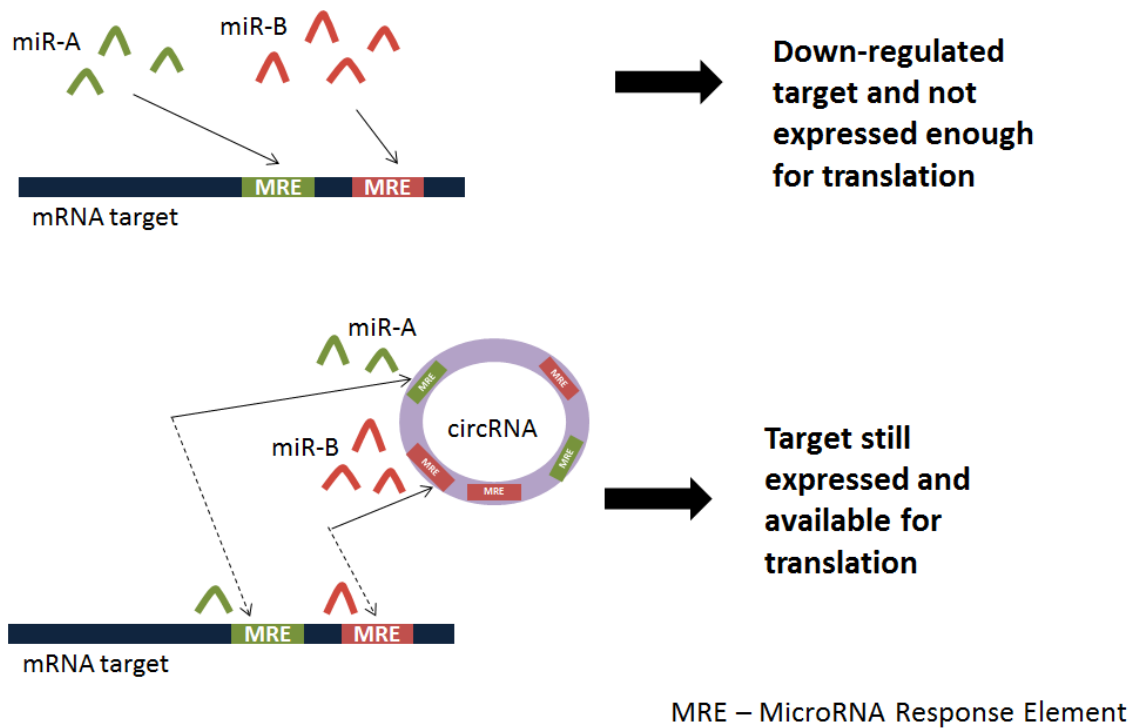
Circular RNAs (circRNAs) are the most recent class of non-coding RNAs which are produced by back-splicing (3' to 5') of precursor mRNAs after transcription. Unlike linear RNAs that contain free 5' and 3' ends, circRNAs have these ends joined together to form a covalently closed loop. Because circRNAs do not have 5' or 3' ends, they are resistant to exonuclease-mediated degradation and are thus expected to be more stable than most linear RNAs in cells. Additionally,

since circRNAs arise from otherwise protein-coding genes, these molecules can also contain MRE sites for microRNA binding. The genomic size of circRNAs can range from few to several kilo-bases and thus can contain a large number of MRE sites. For example, CiRS-7, discovered in 2013 by Memczak et al [16], is a classic example of a circRNA and is a back-spliced product of the CDR1 gene. CiRS-7 was found in human, mouse and nematode brain and consisted of 63 conserved MRE sites for miR-7 binding. Lately, it has also been shown that a subset of circRNAs can be translated into functional proteins [18].

## **1.4 Cross-talk between different RNA types**

The expression of cancer-relevant genes such as tumor suppressors and oncogenes is critical in diseases, such as cancer. Complex interactions between mRNAs, circRNAs, and microRNAs can greatly influence the post-transcriptional activity of such genes in a normal versus cancerous environment within the cell. More precisely, mRNAs and circRNAs contain microRNA binding sites, called MRE sites from hereon, that are complimentary to microRNA seed regions. As shown in Figure 2, mRNAs and circRNAs use their MRE sites to interact with microRNAs. It is known that the interactions between MRE sites on mRNA targets and microRNA seed regions can lead to decreased gene expression or even gene silencing [19, 20]. However, in the presence of circRNAs that share the same MRE sites as mRNA targets, this now exposes microRNAs to non-unidirectional interactions. Such mRNAs and circRNAs can act as competing

endogenous RNAs (ceRNAs) and sequester the same pool of microRNAs using their common MRE sites. As a result, not all microRNAs necessarily bind to the mRNA targets anymore, and thus the gene expression can remain intact.



**Figure 2:** Endogenous competition between mRNAs and circRNAs for a common pool of microRNAs

This form of RNA cross-talk is crucial to understand because the stability of mRNA targets, or lack of stability – depending on how the ceRNAs and microRNAs interact, can cause significant impact to gene expression in a cancerous versus normal environment.

## **1.5 Importance of studying RNA interactions in Breast Cancer**

Breast cancer is the second most common cancer in the United States. Being cancer that is formed in the cells of the breasts, this disease is more common in women. Breast cancer can be non-invasive, invasive or metastatic. Also, depending on the molecular diagnosis, breast cancer can be classified into subtypes based on the hormonal status of three key hormone receptors – estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor (HER2). There are three major molecular subtypes for breast cancer, listed in the order of tumor aggressiveness – ER positive (65 - 80%), HER2 positive (20%) and Triple Negative (10 – 20%).

As mentioned above, among the various molecular subtypes, Triple Negative (TN) cancer is highly heterogeneous and one of the most severe forms of breast cancer. A TN patient is found negative for all three hormone receptors and has a poor prognosis. At present, treatments such as chemotherapy, surgery and radiation therapy or their combination are provided to TN patients; unfortunately, their response to such treatment strategies is indigent, many times with the relapse of the tumor. Due to such treatment failures, there is an incisive need to identify alternative strategies to treat these patients in a more targeted and precise fashion.

On the bright side, we are now in the era of the revolutionary and groundbreaking technology of genome sequencing which is both low-cost and fast, and is famously called the Next Generation Sequencing (NGS) technology. NGS, which became available at the onset of the 21<sup>st</sup> century [21], has now become the standard in bioinformatics for analysis of the genome. Amongst the multitude of sequencing platforms offered by NGS, transcriptome profiling, or RNA-Seq, has been a significant breakthrough [22]. The wealth of information offered by RNA-Seq is so vast that it not only helps uncover numerous features of the protein-coding regions, such as gene expression, single nucleotide variants (SNVs) and gene fusions but also helps mine similar details on expressed but untranslated, non-coding regions of the transcriptome, such as lncRNAs, pseudogenes, and even circRNAs.

## **1.6 Availability of high-throughput NGS datasets**

With the advent of NGS, it was clear that this technology could have an enormous impact on life sciences and would be an invaluable resource for the research community to analyze genomic profiles of heterogeneous diseases such as cancer. Thus the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) collaborated in 2006 and built The Cancer Genome Atlas (TCGA) consortium (<https://cancergenome.nih.gov/>). TCGA is a unique repository of high throughput sequencing data from several

NGS application types, such as DNA-Seq, RNA-Seq, microRNA-Seq, methyl-Seq, etc. made available for many different cancer types. Several TCGA cases also have matched tumor and normal-adjacent tissues from the same patient available for most of the NGS data types, making individualized comparative genomics possible.

While many studies have used TCGA datasets to derive meaningful inferences for different cancers [23-25] successfully, many of these studies have been focused on a specific NGS application type. Very few bioinformatics projects [26-30] have utilized TCGA data from different NGS application types to perform integrated analyses for a specific type of cancer.

As stated in the previous section of this chapter – on the cross-talk between different RNA types, complex interactions between ceRNAs and microRNAs can lead to instability of tumor suppressors, oncogenes, and other relevant target genes, paving the way for tumor growth and progression. Such crosstalk between ceRNAs, i.e., mRNAs and circRNAs with microRNAs, has not yet been studied in breast cancer and is investigated for the first time in this dissertation, using TCGA breast cancer RNA-Seq data. This will be a novel undertaking, and we believe that results from this dissertation will provide a significant contribution towards the current knowledge base of medical research in breast cancer.

## 1.7 Challenges and Motivation

Although several bioinformatics packages have been developed to analyze RNA-Seq data [31-36], these are stand-alone packages that function independently, analyze different aspects of the transcriptome and cannot be plugged together easily. Distinct differences in software prerequisites and methodology of these packages make analysis of multiple features a challenging feat. One of the roadblocks in RNA-Seq data analysis has been the lack of easy integration of such packages as well as the absence of all-inclusive packages or workflows that can stand as a unified solution for interpreting the transcriptome at a broader scale.

Another significant gap found in bioinformatics research is that the depth of information offered by RNA-Seq technology is often not utilized to its full extent. RNA-Seq data can be mined for both protein-coding and non-coding regions using comprehensive and novel bioinformatics techniques; however, this incredible opportunity has often been overlooked. Integrative analyses including both protein-coding and multiple non-coding regions, such as mRNAs and both circRNAs and microRNAs respectively, are essential to enable precise and deeper understanding of the expression, interaction, and regulation of these regions in the transcriptome.

Thus, in light of the above bioinformatics challenges as well as the open possibilities to utilize RNA-Seq data to explore coding and non-coding RNA interactions in cancer patients with both tumor and normal-adjacent samples from same patient, and importantly, with a devoted motivation to strive towards improving diagnostic targets in breast cancer, I hereby present my dissertation with the following objectives:

**1. Development of comprehensive bioinformatics workflows:** develop bioinformatics workflows for the identification, characterization, and quantification of

- a. MRNAs (protein-coding)
- b. CircRNAs (non-coding), and
- c. Integration of mRNAs and microRNAs

**2. Bioinformatics workflows application in TCGA breast cancer:**

Application of the developed bioinformatics workflows to high-throughput TCGA RNA-Seq breast cancer dataset to detect mRNAs and circRNAs

**3. MRE quantification and RNA integration analysis in breast cancer:**

Analysis of TCGA breast RNA-Seq results from the three workflows and TCGA breast microRNA data to

- a. quantify microRNA response element (MRE) sites



- b. identify differentially expressed MRE sites between tumor and normal-adjacent, and
- c. use results from MRE analysis to identify competing endogenous RNA networks in breast tumor that involve mRNAs, circRNAs, and microRNAs

## 1.8 Outline of the chapters

The objectives of my dissertation are formulated to address the challenges as listed in the section above specifically. The outcomes of these objectives are described in the subsequent chapters of this dissertation. A brief outline of these chapters is provided below.

**Chapter 2** introduces the Mayo Analysis Pipeline for RNA-Seq called MAP-RSeq [37]. MAP-RSeq is a comprehensive bioinformatics workflow developed for the identification of mRNAs from RNA-Seq data and can also be used to obtain various genomic features of mRNAs, such as gene and exon expression quantification, single nucleotide variants and gene fusions, for human as well as any well-annotated genome. MAP-RSeq can be used on high-performance computing clusters and can also be run on a single node. The comprehensive reporting style, as well as extensive post-analysis quality checks of MAP-RSeq,

offers an end-to-end solution to researchers for their RNA-Seq bioinformatics needs.

**Chapter 3** describes the bioinformatics workflow developed for the identification and characterization of circRNAs, called Circ-Seq [38]. Circ-Seq is a comprehensive and configurable workflow with unique filters designed to report expressed circRNA candidates. Furthermore, this chapter explains the results obtained from the application of Circ-Seq to a) breast cancer cell lines and b) the biggest cohort of breast cancer samples from the TCGA consortium. Validation experiments using qRT-PCR and Sanger sequencing of a 7kb long circRNA identified in MCF7 breast cancer cell line by Circ-Seq is also a significant part of the results in this chapter.

**Chapter 4** describes a novel bioinformatics approach called ReMlx that utilizes TCGA RNA-Seq data to integrate mRNAs and microRNAs by identifying and quantifying microRNA binding or MRE sites in tumor and normal-adjacent samples of the Triple Negative breast cancer subtype. Differentially expressed MREs are selected, and genes and microRNA candidates associated with these MREs are analyzed in further detail. Significant canonical pathways are identified, and the MRE-associated genes and microRNAs, as well as circRNAs that belong to the genes in these pathways, are studied to identify complex

interactions between mRNAs, circRNAs, and microRNAs in Triple Negative breast cancer.

**Chapter 5** is a conclusion of this dissertation. This chapter offers a quick recap of the bioinformatics challenges faced and the three specific objectives pursued in this dissertation. The results accomplished from individual objectives are outlined. Further, chapter 5 offers future insights into potential directions in which work from this dissertation can be carried forward.

# **Chapter 2: MAP-RSeq – Mayo Analysis Pipeline for RNA Sequencing**

## **2.1 Background**

Next-generation sequencing (NGS) technology breakthroughs have allowed us to define the transcriptome landscape for cancers and other diseases [39]. RNA-Sequencing (RNA-Seq) is information-rich; it enables researchers to investigate a variety of genomic features, such as gene expression, characterization of novel transcripts, alternative splice sites, single nucleotide variants (SNVs), fusion transcripts, long non-coding RNAs, small insertions, and small deletions. Multiple alignment software packages are available for read alignment, quality control methods, gene expression and transcript quantification methods for RNA-Seq [40-43]. However, the majority of the RNA-Seq bioinformatics methods are focused only on the analysis of a few genomic features for downstream analysis [44-47]. At present, there is no comprehensive RNA-Seq workflow that can simply be installed and used for multiple genomic feature analysis. At the Mayo Clinic, we have developed MAP-RSeq - a comprehensive computational workflow, to align, assess and report multiple genomic features from paired-end

RNA-Seq data efficiently with quick turnaround time. We have tested a variety of tools and methods to estimate genomic features from RNA-Seq data accurately. Best performing publically available bioinformatics tools along with parameter optimization were included in our workflow. As needed we have integrated in-house methods or tools to fill in the gaps. We have thoroughly investigated and compared the available tools and have optimized parameters to make the workflow run seamlessly for both the virtual machine and cluster environments. Our software has been tested with paired-end sequencing reads from all Illumina platforms. Thus far, we have processed over 5,000 Mayo Clinic samples using the MAP-RSeq workflow. The MAP-RSeq research reports for RNA-Seq data have enabled Mayo Clinic researchers and clinicians to exchange datasets and findings. Standardizing the workflow has allowed us to build a system that enables us to investigate across multiple studies within the Mayo Clinic. MAP-RSeq is a production application that allows researchers with minimal expertise in LINUX or Windows to install, analyze and interpret RNA-Seq data.

## **2.2 Availability and requirements**

Project name: MAP-RSeq

Project home page: <http://bioinformaticstools.mayo.edu/research/maprseq/>

Operating system(s): Linux or VM

Programming language: PERL, Python, JAVA, R, and BASH

Other requirements: none

License: Open Source

Any restrictions to use by non-academics: none

## **2.3 Implementation**

MAP-RSeq uses a variety of freely available bioinformatics tools along with in-house developed methods using Perl, Python, R, and Java. MAP-RSeq is available in two versions. The first version is single threaded and runs on a virtual machine (VM). The VM version is straightforward to install. The second version is multi-threaded and is designed to run on a cluster environment.

### **2.3.1 Virtual machine**

Virtual machine version of MAP-RSeq is available for download at the provided URL [48]. This includes a sample dataset, references (limited to chromosome 22), and the complete MAP-RSeq workflow pre-installed. Virtual Box software (free for Windows, Mac, and Linux at [49]) needs to be installed on the host system. The system also needs to meet the following requirements: at least 4GB of physical memory, and at least 10GB of available disk. Although our sample data is only from Human Chromosome 22, this virtual machine can be extended to the entire human reference genome or to other species. However, this requires allocating more memory (~16GB) than may be available on a typical desktop system and build the index references files for the species of interest.

Tables 1 and 2 shows the install and runtime metrics of MAP-RSeq in virtual machine and Linux environments respectively. For Table 2, we downloaded the breast cancer cell line data from CGHub [50] and randomly chose 4 million reads to run through the QuickStart VM. It took 6 hours for the MAP-RSeq workflow to complete. It did not exceed the 4GB memory limit but did rely heavily on the swap space provided; making it run slower than if it would have had more physical memory available. Job profiling indicates that the system could have used 11GB of memory for such a sample.

<b>QuickStart VM</b>	<b>File size</b>	<b>Timeline</b>
Download	2.2GB	~ 20 minutes to download on consumer grade internet
Unpacked size	8GB	-
Time to import into VM	-	~ 10 minutes
VM boot	-	3 minutes
Runtime with sample data (chr22 only)	-	~ 30 minutes

**Table 1:** MAP-RSeq installation and runtime for QuickStart virtual machine

<b>Linux</b>	<b>File size</b>	<b>Timeline</b>
Download	930 MB	~10 minutes to download on consumer grade internet
Install time	-	~6 hours (mostly downloading and indexing references)
Unpacked size	9GB	-
Runtime	-	Depends on the sample data used

**Table 2:** MAP-RSeq installation and runtime in a Linux environment

### 2.3.2 Sun grid engine

MAP-RSeq requires four processing cores with a total of 16GB RAM to get optimal performance. It also requires 8GB of storage space for tools and reference file installation. For MAP-RSeq execution the following packages such as JAVA version 1.6.0\_17 or higher, Perl version 5.10.0 or higher, Python version 2.7 or higher, Python-dev, Cython, Numpy and Scipy, gcc and g++ , Zlib, Zlib-devel, ncurses, ncurses-devel, R, libgd2-xpm, and mailx need to be preinstalled and referenced in the environment path. It does also require having additional storage space for analyzing input data and writing output files. MAP-RSeq uses bioinformatics tools such as BEDTools [34], UCSC Blat [51], Bowtie [52], Circos [53], FastQC [54], GATK [33], HTSeq [55], Picard Tools [56], RSeqQC [57], Samtools [58], and TopHat [59]. Our user manual and README files provide detailed information of the dependencies, bioinformatics tools and



parameters for MAP-RSeq. The application requires configuration, such as run, tool, and sample information files, as described in the user manual.

Table 3 shows the processing time of the workflow across different sequencing read depths. Time was recorded from a server with eight quad-core Intel Xeon 2.67 GHz processors and 530 GB of shared memory using Centos 6. For a sample with 1 million reads, MAP-RSeq completes in less than 2 hours. For samples with 150 million to 300 million reads, MAP-RSeq completes in 12-48 hours depending on the hardware used.

<b>MAP-RSeq processing time</b>	<b>Read Counts</b>
118 minutes	1,000,000
82 minutes	500,000
71 minutes	200,000

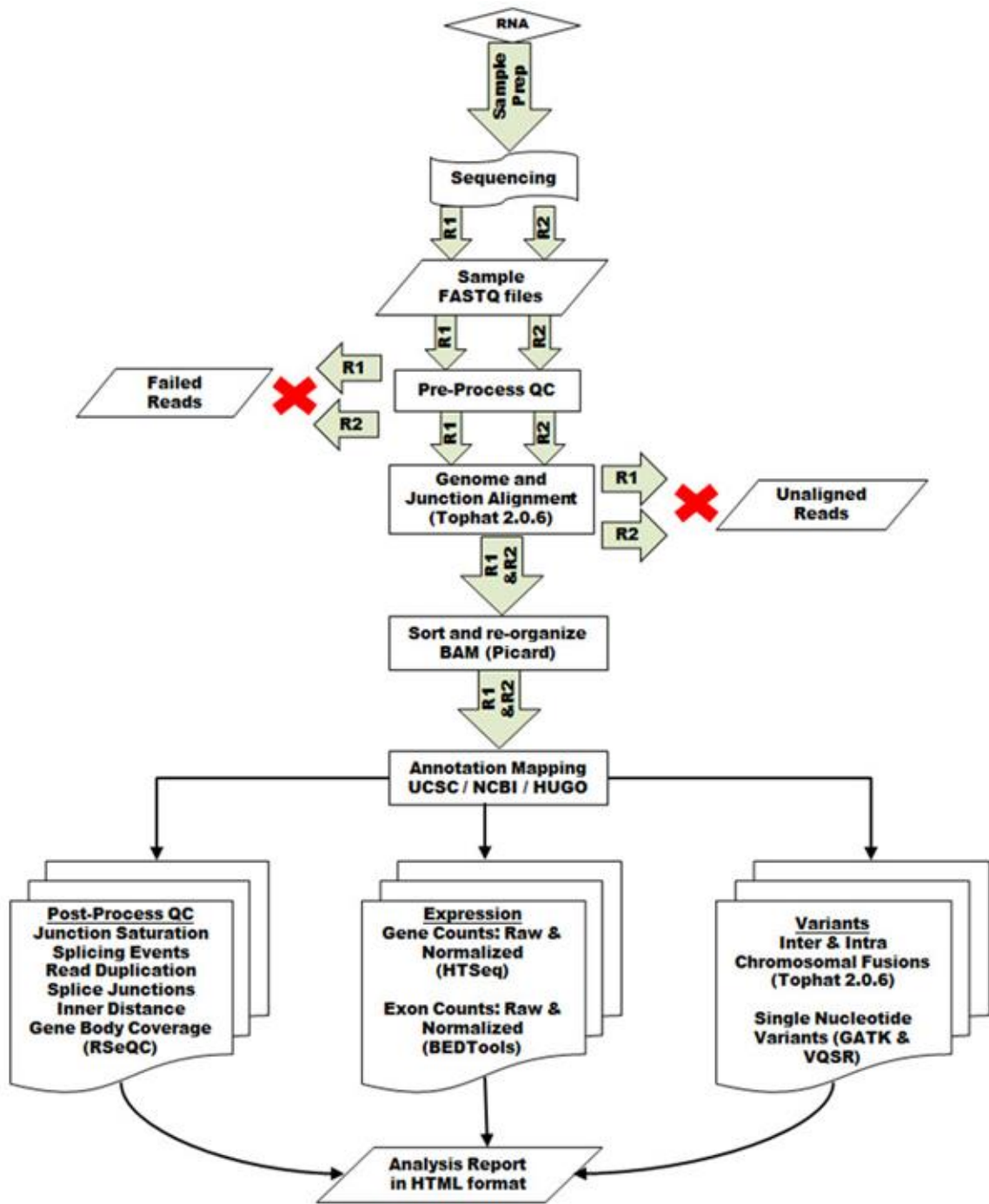
**Table 3:** Wall clock times to run MAP-RSeq at different read counts

## 2.4 Results

NGS technology has been outpacing bioinformatics. MAP-RSeq is a comprehensive simple-to-use solution for analysis of RNA-Sequencing data. We have used both simulated and real datasets to optimize parameters of the tools included in the MAP-RSeq workflow. The high-level design of MAP-RSeq is shown in Figure 3. MAP-RSeq consists of the six major modules such as alignment of reads, quality assessment of sequence reads, gene expression and

exon expression counts, expressed SNVs from RNA-Seq, fusion transcript detection, summarization of data and final report.

Reads are aligned by TopHat 2.0.6 [59] against the human reference genome build (default = hg19) using the bowtie1 aligner option. Bowtie is a fast memory efficient, short sequence aligner [52]. The remaining unaligned reads from Bowtie are used by TopHat to find splice junctions and fusions. At the end of the alignment step, MAP-RSeq generates binary alignment (BAM) and junction bed files for further processing. The workflow uses the RSeQC software [57] to estimate the distance between paired-end reads, evaluate sequencing depth for alternate splicing events, determine the rate of duplicate reads, and calculate coverage of reads across genes as shown in the example report file (Figure 4). The summary statistics and plots generated by MAP-RSeq workflow are used for further quality assessments. The example MAP-RSeq result set (files and summary report) from a RNA-Sequencing run can be downloaded from the MAP-RSeq homepage [48].



**Figure 3:** Flowchart of the MAP-RSeq workflow. High-level representation of the MAP-RSeq workflow for processing RNA-Seq data.

V. Results Summary:

- **QC steps - FastQC-report**  
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.  
FastQC Reports

• **Statistics based on per Sample Analysis (ColumnDescription)**  
Analysis is carried out using fasto sequence files as input and generates output tables. For paired-end runs, the tables contain counts for each sample combined from both reads.

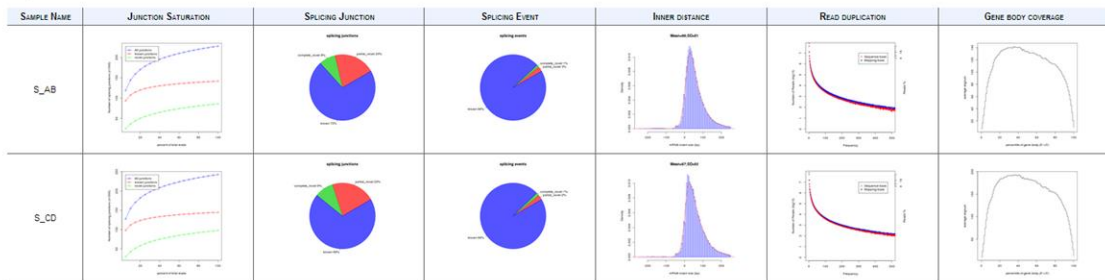
SAMPLE(S)	TOTAL READS	USED READS	MAPPED READS	MAPPED READS (GENOME)	MAPPED READS (JUNCTION)	GENE COUNT	EXON COUNT	SNVs IDENTIFIED
s_AB	294,030,280	282,256,623	282,321,294 (89.2)	236,596,852 (80.5)	25,722,442 (8.7)	163,745,488 (55.7)	185,350,787	292,827
s_CD	367,467,844	366,429,975	350,734,057 (95.4)	316,856,109 (86.2)	34,077,948 (8.3)	195,569,950 (53.2)	236,985,171	383,190

VI. Results Delivered

The following three sets of tables are delivered and column description is available in Appendix.

- **Exon table:** contains counts for the number of times an exon has been detected  
count (raw) = sum of exon read counts  
count (RPKM)
- **Gene table:** contains counts for the number of times a gene copy has been detected  
count (raw) = sum of exon read counts, with an exception that if reads start in different exons of the same gene twice, these are counted only once for the gene
- **SIV reports:** contains Single Nucleotide Variants (SIV) called using GATK software  
sample.gatk.vcf = raw SIV calls for each sample  
sample.filter.vcf = SIV calls annotated using VQSQR filters

[TOP]



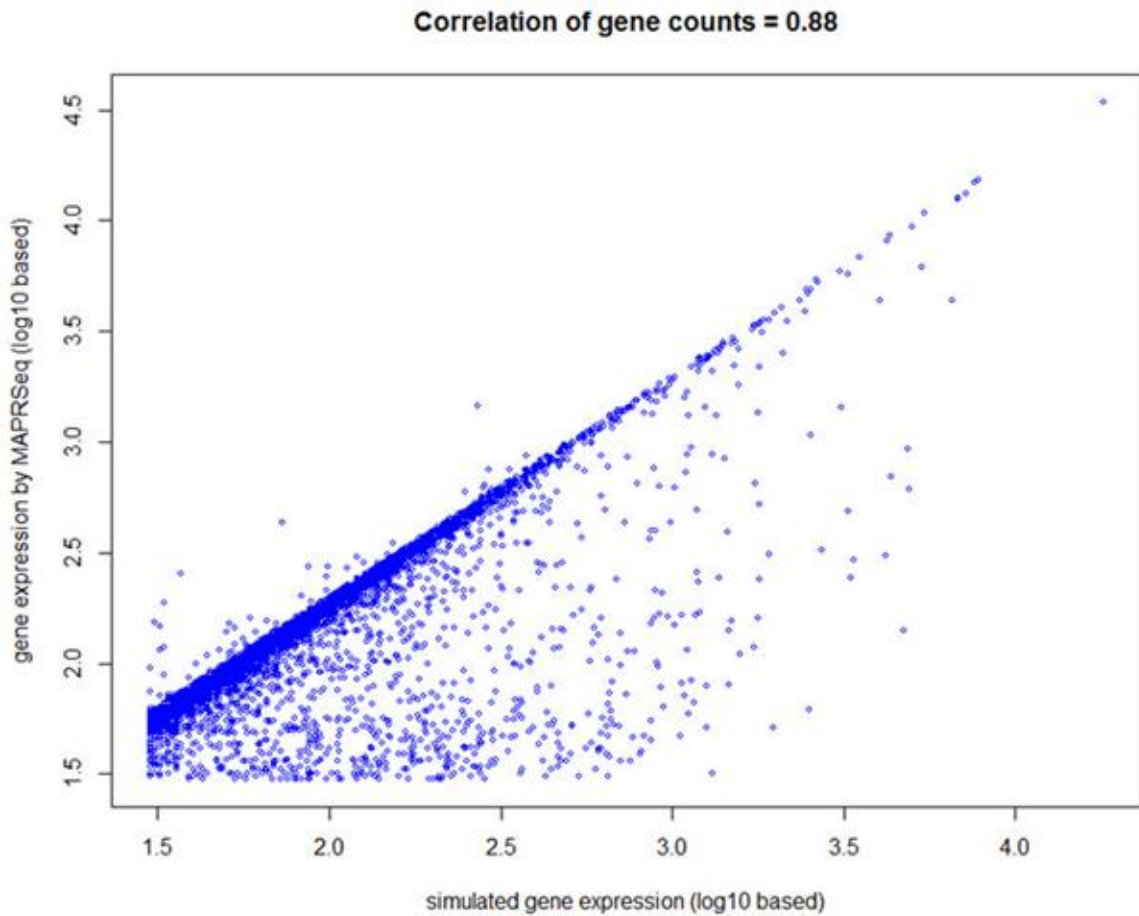
**Figure 4:** Screenshot output report (html) of MAP-RSeq. An example screenshot report of the MAP-RSeq output file.

Several research and clinical projects [60-62] at Mayo Clinic have applied MAP-RSeq workflow for obtaining gene expression, single nucleotide variants and fusion transcripts for a variety of cancer and disease-related studies. Currently, there are multiple ongoing projects or clinical trial studies for which we generate both RNA-Sequencing and exome sequencing datasets at the Mayo Clinic Sequencing Core. We have developed our RNA-Seq and DNA-Seq workflows such that sequencing data can be directly supplied to the pipelines with less manual intervention. Analysis of the next generation sequencing datasets along with phenotype data enables a further understanding of the genomic landscape to better diagnose and treat patients.

### **2.4.1 Gene expression and exon expression read counts**

A Gene expression count is defined as the sum of reads in exons for the gene whereas an exon expression count is defined as the sum of reads in a particular exon of a gene. Gene expression counts in MAP-RSeq pipeline can be obtained using HTSeq [55] software (default) or featureCounts [32] software. The gene annotation files were obtained from the Cufflinks website [63]. Exon expression counts are obtained using the intersectBed function from the BEDTools Suite [34].

MAP-RSeq gene expression counts module was validated using a synthetic dataset for which RNA-Seq reads were simulated using the BEERS software - a computational method that generates paired-end RNA-sequencing reads for Illumina platform [64]. The parameters used for BEERS to generate simulated data are: total reads = 2 million reads, hg19 annotation from RefSeq, read length = 50 bases, base error = 0.005 and substitution rate = 0.0001. Simulated reads were aligned and mapped using the MAP-RSeq workflow. The mapped reads were then input into HTSeq for gene expression counts. Genes with fewer than 30 reads were excluded from the analysis. A correlation of  $r = 0.87$  was observed between the Reads Per Kilobase per Million (RPKM) simulated gene counts and the counts reported by MAP-RSeq, as shown in Figure 5.



**Figure 5:** Correlation of gene counts reported by MAP-RSeq in comparison to counts simulated by BEERS. MAP-RSeq uses the HTSeq software to classify reads to genomic features. The intersection nonempty mode of HTSeq was applied, and the query-name sorted alignment (BAM) file along with the reference GTF file obtained from BEERS were provided as input files to HTSeq for accurate assignment of paired-end reads to genomic features. Comparison of the gene counts (RPKM) obtained from MAP-RSeq with counts for respective genes simulated by BEERS yielded a Pearson correlation of 0.87. The genomic regions where gene expression reported by HTSeq did not completely correlate with simulated expression are due to ambiguous reads or due to the fact that either mate of the paired-end read mapped to a different genomic feature, thus categorizing the read as ambiguous by HTSeq.

For simulated data (50 bases), Table 4 summarizes various statistics reported by the MAP-RSeq workflow regarding the alignment of reads to transcriptome and

junctions, gene and exon abundance as well as number of SNVs identified and annotated using GATK.

<b>MAP-RSeq features</b>	<b>Statistics</b>
Total number of single reads	4,000,000
Reads used for alignment	3,999,995
Total number of reads mapped	3,851,539 (96.3%)
Reads mapped to transcriptome	3,401,468 (85.0%)
Reads mapped to junctions	450,071 (11.3%)
Reads contributing to gene abundance	1,395,844
Reads contributing to exon abundance	11,266,392
Number of SNVs identified	6,222

**Table 4:** Alignment statistics from MAP-RSeq using a simulated dataset from BEERS

An example of MAP-RSeq gene counts table; exon counts table, and normalized counts (RPKM) along with annotations for each run are shown in Figure 6.

Chr	GeneID	Start	Stop	CodingLength	s AB GeneCount	s AB RPKM	s CD GeneCount	s CD RPKM
chr1	AADACL3	12776118	12788726	4049	0	0	0	0
chr1	AADACL4	12704566	12727097	1575	0	0	0	0
chr1	ABCA4	94458394	94586705	7325	6	0.003122555	4	0.001556949
chr1	ABCB10	229652329	229694442	3857	2180	2.154633008	3150	2.328536104
chr1	ABCD3	94883933	94984219	3797	1658	1.664601889	2278	1.710547678
chr1	ABL2	179068462	179198819	12649	4442	1.338717115	6520	1.469648461
chr1	ACADM	76190043	76229355	2615	524	0.763881598	544	0.593129149
chr1	ACAP3	1227764	1243269	3759	8496	8.616058362	11564	8.771175857
chr1	ACBD3	226332380	226374423	3565	7540	8.662658387	10676	8.53829375
chr1	ACBD6	180257352	180472022	1616	1760	3.208218996	1554	2.741774413
chr1	ACOT11	55013807	55100417	3391	84	0.094431731	140	0.117712425
chr1	ACOT7	6324332	6453826	8389	412	0.657426976	546	0.651626225
chr1	ACP6	147119168	147142634	1808	566	1.193395674	374	0.589787056
chr1	ACTA1	229566993	229603843	1492	94	0.24017372	54	0.10319223
chr1	ACTL8	18081888	18153558	1861	0	0	0	0
chr1	ACTN2	236849770	236927558	4528	2	0.001683798	4	0.002518695
chr1	ACTRT2	2938046	2939467	1422	0	0	0	0
chr1	ADAM15	155023762	155035252	2967	8386	10.77466474	12116	11.64297003
chr1	ADAM30	120436156	120439147	2992	2	0.002548208	8	0.007623431
chr1	ADAMTS4	161159538	161168845	4332	678	0.59663359	910	0.598928536
chr1	ADAMTSL4	150521898	150533412	4299	22900	20.30647268	36388	24.13308275
chr1	ADAR	154554534	154600456	7092	95346	51.25074763	203616	81.85877398
chr1	ADC	33546714	33585995	2182	146	0.255073043	154	0.201227826
chr1	ADCK3	227127938	227175246	2924	10182	13.27462246	8164	7.960634583

Chr	Start	Stop	Gene	s AB ExonCount	s AB RPKM	s CD ExonCount	s CD RPKM
chr1	11874	12227	DDX11L1	0	0	0	0
chr1	12613	12721	DDX11L1	0	0	0	0
chr1	13221	14408	DDX11L1	3	0.009626563	6	0.014399814
chr1	14362	14829	WASH7P	66	0.537606532	79	0.481286078
chr1	14970	15038	WASH7P	6	0.331488612	12	0.495854452
chr1	15796	15947	WASH7P	17	0.426355419	19	0.356395387
chr1	16607	16765	WASH7P	3	0.071926774	7	0.125522904
chr1	16858	17055	WASH7P	5	0.096265632	13	0.187197577
chr1	17233	17368	WASH7P	0	0	0	0
chr1	17606	17742	WASH7P	0	0	0	0
chr1	17915	18061	WASH7P	0	0	1	0.019395667
chr1	18268	18366	WASH7P	0	0	1	0.028799627
chr1	24738	24891	WASH7P	1	0.02475402	0	0
chr1	29321	29370	WASH7P	1	0.076242381	0	0
chr1	34611	35174	FAM138A	1	0.006759076	0	0
chr1	34611	35174	FAM138F	1	0.006759076	0	0
chr1	35277	35481	FAM138A	1	0.018595703	0	0
chr1	35277	35481	FAM138F	1	0.018595703	0	0
chr1	35721	36081	FAM138A	1	0.010559887	0	0
chr1	35721	36081	FAM138F	1	0.010559887	0	0
chr1	69091	70008	OR4F5	1	0.004152635	0	0
chr1	134773	139696	LOC729737	6090	4.714826354	7659	4.434820909
chr1	139790	139847	LOC729737	952	62.57133324	1099	54.02462488

**Figure 6:** Screenshots of gene and exon expression reports by MAP-RSeq. An example of the gene and exon expression counts from the output reports of MAP-RSeq.



## **2.4.2 Differential expression**

Each sample is associated with a phenotype, such as a tumor, normal, treated, control, etc and that meta-data needs to be obtained to form groups for differential expression analysis. To remove any outlier samples, it is required to perform detailed quality control checks prior to gene expression analysis. There are a variety of software packages that are used for differential expression analysis using RNA-Seq gene expression data [42, 65-67]. Several studies have been published comparing the differential expression methods and concluded that there are substantial differences regarding sensitivity and specificity among the methods [68-70]. We have chosen edgeR software [42] from R statistical package for gene expression analysis. In our source code for MAP-RSeq pipeline, we have Perl, R scripts and instructions that can be used post MAP-RSeq run for differential expression analysis.

## **2.4.3 Expressed SNVs (eSNVs) from RNA-Seq**

After filtering out multiple mapped and fusion reads, the MAP-RSeq calls SNVs using UnifiedGenotyper v.1.6.7 and VariantRecalibrator from Genome Analysis ToolKit (GATK) with the alignment files generated by Tophat. The UnifiedGenotyper from GATK is a single nucleotide variant (SNV) and indel caller developed by the BROAD institute [33]. SNVs are further annotated by the variant quality score recalibration (VQSR) method. The annotated SNVs are further filtered based on read quality (QD), coverage (DP), strand bias (FS), and

positional bias (ReadPosRankSum) to identify true variants. A 1000 genome sample (NA07347) with both exome and RNA-Seq data was used to validate the SNV calling module of MAP-RSeq workflow. A concordance rate of 95.6% was observed between the MAP-RSeq SNV calls, and the exome sequencing variant calls for NA07347. Figure 7 shows a screenshot of the MAP-RSeq variant calling file. Confident variant calls from MAP-RSeq workflow at high and low read depths of sequencing are shown in Figure 8A and 8B respectively.

```

##fileformat=VCFv4.1
##FILTER=ID=DFilter;Description="ED > 5"
##FILTER=ID=FSFilter;Description="FS > 20.0"
##FILTER=ID=RFSEFilter;Description="ReadPosRankSum > 8.0"
##FORMAT=ID=AN;Number=1;Type=Integer;Description="Allelic depths for the ref and alt alleles in the order listed"
##FORMAT=ID=DP;Number=1;Type=Integer;Description="Approximate read depth (reads with MQ<255 or with bad mates are filtered)"
##FORMAT=ID=DQA;Number=1;Type=Integer;Description="The number of high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases"
##FORMAT=ID=CG;Number=1;Type=Float;Description="Genotype Quality"
##FORMAT=ID=GT;Number=1;Type=String;Description="Genotype"
##FORMAT=ID=PL;Number=0;Type=Integer;Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification"
##INFO=ID=AC;Number=A;Type=Integer;Description="Allele count in genotypes, for each ALT allele, in the same order as listed"
##INFO=ID=AF;Number=A;Type=Float;Description="Allele Frequency, for each ALT allele, in the same order as listed"
##INFO=ID=AN;Number=1;Type=Integer;Description="Total number of alleles in called genotypes"
##INFO=ID=BaseQRankSum;Number=1;Type=Float;Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref base qualities"
##INFO=ID=DB;Number=0;Type=Flag;Description="dBSNP Membership"
##INFO=ID=DP;Number=1;Type=Integer;Description="Approximate read depth, some reads may have been filtered"
##INFO=ID=DS;Number=0;Type=Flag;Description="Were any of the samples downsampled?"
##INFO=ID=Deln;Number=1;Type=Float;Description="Fraction of Reads Containing Splicing Deletions"
##INFO=ID=ED;Number=1;Type=Integer;Description="Number of blat hits to reference genome, not counting self-hit"
##INFO=ID=FS;Number=1;Type=Float;Description="Phred-scaled p-value using Fisher's exact test to detect strand bias"
##INFO=ID=HRun;Number=1;Type=Integer;Description="Largest Contiguous Homopolymer Run of Variant Allele in Either Direction"
##INFO=ID=HaplotypeScore;Number=1;Type=Float;Description="Consistency of the site with at most two segregating haplotypes"
##INFO=ID=InbreedingCoeff;Number=1;Type=Float;Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation"
##INFO=ID=MQ;Number=1;Type=Integer;Description="MQ Mapping Quality"
##INFO=ID=MQ0;Number=1;Type=Integer;Description="Total Mapping Quality Zero Reads"
##INFO=ID=MQRankSum;Number=1;Type=Float;Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read mapping qualities"
##INFO=ID=QD;Number=1;Type=Float;Description="Variant Confidence/Quality by Depth"
##INFO=ID=ReadPosRankSum;Number=1;Type=Float;Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias"
##INFO=ID=SB;Number=1;Type=Flag;Description="Strand Bias"
##initialGenotype="analysis_type=UnionGenotype read_buffer_size=null phone_home=NO_ET read_filter={} excludeIntervals=null interval_set_rule=UNION interval_merging=ALL nonDeterministicRandomSeed=false
downsampling_type=BY_SAMPLE downsampling_to_fraction=null downsampling_to_coverage=250 baq=OFF baqOpenPenalty=40 0 performanceLog=null useOriginalQualities=false BQSR=null quantize_qual=1
defaultBaseQualities=1 validation_strictness=SILENT unseterrnull num_threads=null num_cpu_threads=null num_io_threads=null num_bwa_file_handles=null read_group_black_list=null pedigreeString=[]
pedigreeValidationType=STRICT allow_intervals_with_unindexed_bam=false logging_level=INFO log_to_file=null help=false genotype_likelihoods_model=SNP p_score=1 heterozygosity=0.010 pcr_error_rate=
1.0E-4 genotyping_mode=DISCOVER output_mode=EMIT VARIANTS_ONLY standard_min_confidence_threshold_for_calling=30 0 standard_min_confidence_threshold_for_mating=30 0 noSD=false annotateMD=false alleles=
{[RoBinding name=source=UNBOUND] min_base_quality_score=17 max_deletion_fraction=0.05 max_alternate_allele=1 min_indel_count_for_genotyping=5 min_indel_fraction_per_sample=0.25 indel_heterozygosity=1.25E-4
indelGapContractionPenalty=19 indelGapOpenPenalty=45 indelHaplotypeSize=40 noIndexedIndel=false indelDebug=false ignoreSNPs=false dmp=(RoBinding name=dmp source=
/data2/b1/reference/annotation/dBSNP/hg19/dbsnp_135_hg19.vcf.gz) comp=[] out-org broadinstitute stng gatk io stubs VCFWriterStub NO_HEADER=org broadinstitute stng gatk io stubs VCFWriterStub sites_only=
org broadinstitute stng gatk io stubs VCFWriterStub debug_file=null metrics_file=null annotation={} excludeAnnotation={} filter_mismatching_base_and_qual=false
nonDeterministicRandomSeed=false downsampling_type=BY_SAMPLE downsampling_to_fraction=null downsampling_to_coverage=1000 baq=OFF baqOpenPenalty=40 0 performanceLog=null useOriginalQualities=false BQSR=null
quantize_qual=1 defaultBaseQualities=1 validation_strictness=SILENT unseterrnull num_threads=null num_cpu_threads=null num_io_threads=null num_bwa_file_handles=null read_group_black_list=null pedigree=[]
pedigreeString=[] pedigreeValidationType=STRICT allow_intervals_with_unindexed_bam=false logging_level=INFO log_to_file=null help=false variant={[RoBinding name=variant] mask={[RoBinding name=source=UNBOUND]
out-org broadinstitute stng gatk io stubs VCFWriterStub NO_HEADER=org broadinstitute stng gatk io stubs VCFWriterStub sites_only=org broadinstitute stng gatk io stubs VCFWriterStub filterExpression=[FS >
20.0 ED > 5 ReadPosRankSum < 8.0 ReadPosRankSum > 8.0] clusterSize=3 clusterIndovSize=0 maskExtension=0 maskName=Mask missingValuesInExpressionsShouldDeactivateFailing=false invalidatePreviousFilters=
false filter_mismatching_base_and_qual=false
##contig=ID=chr1;length=24895621
##contig=ID=chr10;length=13534747
##contig=ID=chr11;length=135005169
##contig=ID=chr12;length=133851895
##contig=ID=chr13;length=115163878
##contig=ID=chr14;length=107343540
##contig=ID=chr15;length=102531393
##contig=ID=chr16;length=90354753
##contig=ID=chr17;length=81352103
##contig=ID=chr18;length=78977248
##contig=ID=chr19;length=59128983
##contig=ID=chr2;length=41319373
##contig=ID=chr20;length=63025520
##contig=ID=chr21;length=48129395
##contig=ID=chr22;length=51304586
##contig=ID=chr23;length=58022430
##contig=ID=chr24;length=19154276
##contig=ID=chr5;length=180915268
##contig=ID=chr6;length=17115287
##contig=ID=chr7;length=159138663
##contig=ID=chr8;length=165482022
##contig=ID=chr9;length=141213431
##contig=ID=chrX;length=15559
##contig=ID=chrY;length=59235569
##CHROM POS ID REF ALT QUAL FILTER INFO FORMAT s_AB
chr1 14930 r75454227 A C 61.83 DFFilter AC=1:AF=0.58:AN=2:BaseQRankSum=0.322:DB:DP=10:Deis=0.00:ED=6:FS=0.000:HRun=0:HaplotypeScore=0.0000:MQ=50.00:MQ0=0:MQRankSum=0.322:OD=
6.18:ReadPosRankSum=0.322:SB=0.01
chr1 700307 C A 185.44 DFSEFilter AC=1:AF=0.58:AN=2:BaseQRankSum=0.000:DP=24:Deis=0.00:ED=13:FS=3.349:HRun=2:HaplotypeScore=1.9970:MQ=50.00:MQ0=0:MQRankSum=-1.016:OD=
6.89:ReadPosRankSum=1.31:SB=-35.22
chr1 700371 G A 238.53 DFSEFilter AC=1:AF=0.58:AN=2:BaseQRankSum=0.296:DP=11:Deis=0.00:ED=86:FS=3.233:HRun=0:HaplotypeScore=0.9665:MQ=50.00:MQ0=0:MQRankSum=-1.580:OD=
20.78:ReadPosRankSum=1.386:SB=-123.70
GT:AD:DP:RF:GQ:PI
0/1:3:7:11:0:3:3:4:93:59:259:0:94

```

Figure 7: Screenshot of a MAP-RSeq VCF file after VQSR annotation. An example of SNV data representation from MAP-RSeq runs.



#### **2.4.4 Fusion transcript detection**

The TopHat-Fusion algorithm identifies fusion transcripts accurately [31]. MAP-RSeq uses the TopHat-Fusion algorithm and provides a list of expressed fusion transcripts. In addition to the output from TopHat-Fusion, we have implemented modules to visualize fusion transcripts using circos plots [53]. Fusion transcript candidates are reported and summarized by MAP-RSeq. As shown in Figure 9, intra and inter fusion transcripts along with annotations are provided for each sample by the workflow. A circos plot is generated to visualize fusion transcripts across an entire RNA-Seq run. MAP-RSeq also generates 5'–3' fusion spanning sequence for PCR validation of fusion transcripts identified. These primer sequences can be selected by researchers to validate the fusion transcripts.



### **2.4.5 Summarization of data and final report**

The workflow generates two main reports for end users: 1) summary report for all samples in a run with links to detailed reports and six QC visualizations per sample 2) final data report folder consisted of exon, gene, fusion and expressed SNV files with annotations for further statistical and bioinformatics analysis.

A screenshot of an example report from MAP-RSeq is shown in Figure 4.

Detailed descriptions of the samples processed by MAP-RSeq along with the study design and experiment details are reported by the workflow. Results are summarized for each sample in the report. Detailed quality control information, links to gene expression counts, exon counts, variant files, fusion transcript information and various visualization plots are also reported.

## **2.5 Conclusions**

MAP-RSeq is a comprehensive simple-to-use application. MAP-RSeq reports alignment statistics, in-depth quality control statistics, gene counts, exon counts, fusion transcripts, and SNVs per sample. The output from the workflow can be plugged into other software or packages for subsequent downstream bioinformatics analysis. Several research and clinical projects at the Mayo Clinic have used the gene expression, SNVs and fusion transcripts report from the MAP-RSeq workflow for a wide range of cancers and other disease-related

studies. In future, we plan to extend our workflow such that alternate splicing transcripts and non-coding RNAs can also be obtained.

# **Chapter 3: CircRNAs and their associations with breast cancer subtypes**

## **3.1 Introduction**

Circular RNAs (circRNAs) are recently discovered members of noncoding RNAs. They range in length from a few hundred to thousands of nucleotides [71]. In contrast to linear RNA transcripts, which are normally spliced tail-to-head, circRNAs are formed by the covalent bonding of their 3' and 5' (head-to-tail) ends [72]. The lack of open sites at the 5' and 3' ends exempts circRNAs from exonuclease degradation [73], making them stable in cells [74]. When circRNAs were initially identified in plants, they were considered pathogenic because of their structural similarity to viruses [75, 76]. Meanwhile, circRNAs observed in mammalian cells around the same time were thought to result from splicing errors [77-79]. However, more recent studies of circRNAs in *Drosophila*, mouse, and other eukaryotes suggest that these RNA molecules are evolutionarily conserved and thus are not simple artifacts of faulty splicing [80, 81]. In addition, advances in sequencing technology and bioinformatics analyses have renewed interest in these forms of RNA transcripts [16, 72, 82].



After discovering that circRNAs are highly abundant in not only *C. elegans* and zebrafish, but also mouse and human, researchers have begun to uncover many intriguing facets about these diverse RNAs [73]. Many studies have confirmed that circRNAs possess significant pre- and post-transcriptional regulatory functions in mammalian cells [71, 82, 83] and changes in the abundance of circRNAs can adversely affect gene expression [84, 85]. Recent studies indicate that some of the most common functions of circRNAs include their active participation in pre-mRNA splicing [80] as well as promoting transcription of their parent mRNAs [86]. Apart from the above, circRNAs can sometimes serve as microRNA sponges, such as the human circRNA CDR1as, which was shown to contain over 70 binding sites for miR-7 [16, 87].

Stable, cell-free circRNAs have been found in saliva [88] and exosomes [89], making them promising candidates for diagnosis and therapeutics. In particular, discovering disease-specific circRNAs could help identify diagnostic targets in heterogeneous diseases such as cancer. Memczak et al. and Salzman et al. have developed bioinformatics approaches to detect circRNAs using high-throughput transcriptome sequencing, and to date, several hundred human circRNAs have been identified and cataloged [16, 72, 90]. However, the significance of these RNAs in health and disease is still poorly understood. Recently, Bachmayr-Heyda et al. reported that colorectal tumor samples have a lower number of circRNAs compared to matched normal colon mucosa [91]. It is

known that circRNAs are also associated with single nucleotide polymorphisms linked to a wide range of diseases, including various types of cancer, Parkinson's disease, Alzheimer's disease, multiple sclerosis, and schizophrenia [92].

Here, we have enhanced existing methodologies of circRNA detection [16] and developed a parallelized and configurable workflow, Circ-Seq, that annotates and reports expressed and exclusive circRNAs as final candidates from the analysis. We applied Circ-Seq to one of the largest transcriptome sequencing data available for breast cancer samples, provided by The Cancer Genome Atlas (TCGA) consortium. We identified unique and novel circRNAs present in breast tumor samples and normal-adjacent breast tissue. We identified circRNAs specific to breast tumor samples and cataloged circRNAs unique to each of the three breast cancer subtypes: triple negative (TN), estrogen receptor positive (ER+), and ErbB2 overexpressed–HER2 positive (HER2+). Notably, a lower number of circRNAs were observed in breast tumors compared to both normal-adjacent breast tissues from TCGA as well as normal mammary tissue samples from GTEx. Finally, using a panel of 11 cell proliferation gene markers (ROR-P score), we show that the number of circRNAs detected in the ER+ tumor is associated with gene proliferation markers [93]. We also demonstrate that Luminal B tumors have a distinct trend compared to Luminal A tumors based on their circRNA numbers. On the basis of its ability to detect circRNAs in breast

cancer samples, we believe that Circ-Seq will be a valuable tool for researchers to identify circRNAs for diagnosis and treatment of complex diseases.

## **3.2 Results**

### **3.2.1 Circ-Seq: an automated workflow for circRNA identification**

Using existing bioinformatics approaches for circRNA identification by Memczak et al. [16], we developed an integrated analytical workflow called Circ-Seq, for identifying and characterizing circRNAs using high-throughput transcriptome sequencing data. Briefly, it improves the existing methodology by applying filters namely, expression, genomic size, and validation filters, to report a more confident and final catalog of expressed candidate circRNAs. The expression filter retains circRNAs based on the desired number of junction-spanning reads, which is configurable based on sequencing throughput of the sample being analyzed. Next, the genomic size filter is applied to discard any circRNA candidate with tail-to-head genomic distance less than six bases. Finally, the validation filter uses BLAT [51] to query circRNAs to ensure they do not represent repetitive regions of the genome. Towards the end of the workflow, circRNA fused junctions of the final candidates are annotated with valuable genomic information. Annotation of whether the circRNA is a spliced product of a single gene ('intra-gene') or formed across 2 or more genes ('inter-gene'), and exon location of its 3' and 5' ends ('exon-exon boundary' or 'within\_exon') are provided for users discretion to prioritize circRNA candidates in the final report.

The workflow is fully automated and designed to run in a multi-threaded cluster environment and can also be used to analyze single-end or paired-end transcriptome samples. Circ-Seq workflow can be downloaded from (<http://bioinformaticstools.mayo.edu/research/circ-seq/>).

### **3.2.2 Identification of circRNAs in breast cancer cell lines**

To demonstrate the utility of Circ-Seq, we first tested the workflow on the transcriptomes of eight cell lines – seven from breast tumors (BT20, BT474, MCF7, MDAMB231, MDAMB468, T47D, and ZR751) and one from non-tumor breast cell line (MCF10A) [94], and validated one of the largest circRNA candidates reported by the workflow.

CircRNAs were expressed in both the tumor and normal breast cell lines. As shown in Table 5, we identified an average of 10 circRNAs in the triple negative (TN) cancer cell lines, 22 in the estrogen receptor positive (ER+) cancer cell lines, and 9 in the non-tumor MCF10A cell line. On average, circRNAs detected in the cancer cell lines had 12 junction supporting reads in both TN and ER+ subtypes. Assuming that the exon-intron structures of circRNAs remain intact [86], we observed variable genomic sizes for circRNAs in the tumor and non-tumor cell lines. While the smallest circRNA was of size 51 bases in the tumor (ZR751 and BT474) and 70 bases in the non-tumor MCF10A, large circRNAs with genomic sizes exceeding 5kb were found in MCF7, BT474, ZR751 and

MDAMB231 tumor lines. After annotating the head-to-tail fused junctions of these circRNAs with gene models, we found that 31% of circRNAs are spliced products of a single gene (intra-gene) and 12% are inter-gene circRNAs. Additionally, 25% of the circRNAs have their head-to-tail fused junctions along legitimate exon-exon boundaries whereas 18% were found with circRNA junctions inside exons and not on the exon boundaries.

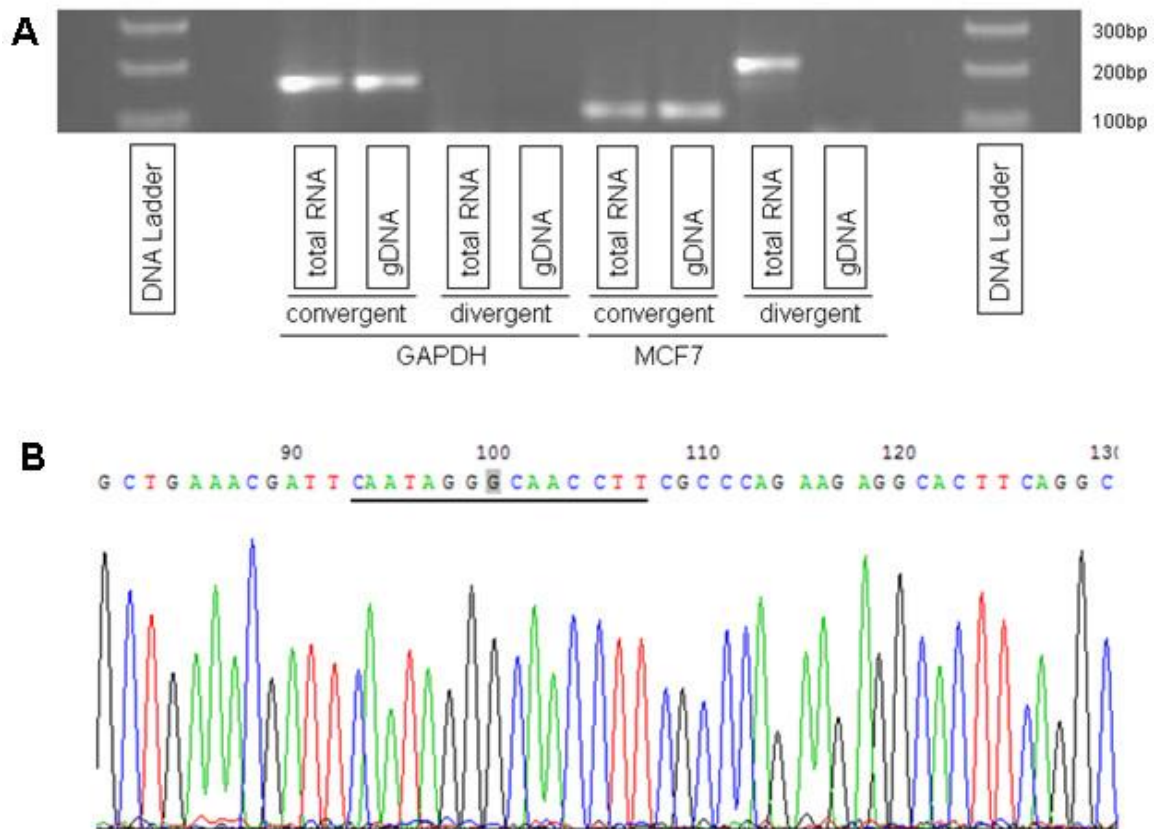
Cell line	Tissue	Breast Cancer Subtype	Total number of circRNAs identified	Final number of circRNAs (after three filters)	Average number of circRNA junction supporting reads
MDAMB231	Tumor	TN	1,111	10	11.2
MDAMB468	Tumor	TN	2,540	15	9.8
BT20	Tumor	TN	1,592	6	15
BT474	Tumor	ER+	4,662	43	14.5
ZR751	Tumor	ER+	3,148	31	11.1
T47D	Tumor	ER+	1,306	5	13.2
MCF7	Tumor	ER+	1,838	9	10
MCF10A	Non-Tumor	–	1,363	9	10.4

**Table 5:** Number of circRNAs identified in breast cell lines using the Circ-Seq workflow

### 3.2.3 Validation of circRNA in MCF7 breast cancer cells

To establish the reliability of circRNA candidates reported by Circ-Seq, we validated one of the largest circRNAs identified in MCF7, the most widely accessible tumor breast cell line that was available in-house. Circ-Seq results for MCF7 indicated that 2 out of 9 circRNAs were found to span legitimate exon-exon boundaries, of which one had a genomic size of 64 bases and the other 7

kb. Since some circRNAs were previously reported to act as microRNA sponges and thus had to be long enough to harbor multiple microRNA binding sites [87], we decided to select the largest out of the two circRNAs in MCF7 for validation. This circRNA was found at chr14:102,466,325–102,500,789 and had 12 supporting junction-spanning reads. The validation consisted of using two independent sets of qRT-PCR experiments. To validate the existence of circRNAs, two different primers were prepared – convergent and divergent [16]. Convergent primers are traditional primers that confirm the existence of linear or tail-to-head (5' to 3') RNA transcripts, however divergent primers are designed in a circular or head-to-tail fashion (3' to 5') to enable binding to circRNA fragments for validation. As shown in Figure 10A, the divergent primer amplified circRNA from MCF7 total RNA but not from genomic DNA (gDNA) whereas GAPDH, which was used as a control, had no results from divergent primers but confirmed its linear RNA using convergent primers. Additionally, Sanger sequencing of the qRT-PCR product validated the head-to-tail splicing. In Figure 10B, the underlined genomic sequence CAATAGGGCAACCTT represents the circRNA spliced junction with the 3' tail fusing to 5' head at the highlighted 'G' nucleotide.



**Figure 10:** Validation of a circRNA at locus chr14:102,466,325–102,500,789. (A) circRNA was amplified by divergent primers using total RNA but not genomic DNA (gDNA). GAPDH was used as a control. (B) Head-to-tail splicing was confirmed by Sanger Sequencing.

### 3.2.4 Presence of circRNAs in TCGA breast cancer transcriptomes

We applied Circ-Seq workflow to 885 whole-transcriptome sequences from breast tumor and normal-adjacent samples provided by the TCGA consortium. Our goal was to use this unique repository to identify circRNAs that differ

between normal-adjacent and tumor tissue. CircRNA results from the workflow for 885 RNA-Seq breast TCGA samples are available for download at <https://noncodingrnaexplorer.wordpress.com>.

### **3.2.4.1 Breast cancer subtype analysis**

#### **3.2.4.1.1 *circRNAs in tumors and normal-adjacent tissue***

Using the Circ-Seq workflow, we processed 128 tumor and 13 normal-adjacent TN samples, 503 tumor and 56 normal-adjacent ER+ samples, and 162 tumor and 20 normal-adjacent HER2+ samples. As shown in Table 6, we observed a total number of 4,542 and 342 circRNAs in tumor and normal-adjacent samples respectively for the TN subtype. Next, we found the number of unique circRNAs that represented exclusive genomic coordinates in tumor and normal-adjacent samples. Note that a unique circRNA is counted once although it may occur in 2 or more samples with the same genomic coordinate. We observed 1,395 unique circRNAs in TN tumor samples and 208 circRNAs in normal-adjacent tissue samples. Similarly, we identified 14,113 (total) and 3,012 (unique) circRNAs in ER+ tumor samples and 2,317 (total) and 1,409 (unique) circRNAs in normal-adjacent tissue samples. Finally, 6,340 (total) and 2,660 (unique) circRNAs were identified in HER2+ tumors and 532 (total) and 284 (unique) in normal-adjacent tissue samples. A summary of the unique circRNAs for the three breast cancer subtypes is shown in Table 6.



Categories	Triple Negative (TN)		Estrogen Receptor (ER+)		ERBB2 overexpressed (HER2+)	
	Tumor	Adjacent	Tumor	Adjacent	Tumor	Adjacent
Total number of samples	128	13	503	56	162	20
Total number of circRNAs	4,542	342	14,113	2,317	6,340	532
Total number of unique circRNAs	1,395	208	3,012	1,409	2,660	284
Ratio of total circRNAs to samples	35	26	28	41	39	27
Ratio of unique circRNAs to samples	12	16	7	25	17	14
Number of unique circRNAs seen in 10% or more samples	729	162	1,086	455	896	193
Number of tumor-specific circRNAs	256	–	288	–	411	–

**Table 6:** Summary of breast tumors, adjacent tissues, and tumor-specific circRNAs in sequence data made available by the cancer genome atlas

We further investigated the unique circRNAs between tumor and normal-adjacent to find circRNAs distinct to the tumor. We observed that normal circRNAs spanned larger genomic regions (from 3' head to 5' tail). Interestingly, within the same genomic region of the normal circRNAs, we found one or more smaller circRNAs that belonged to the tumor samples. Assuming that circRNAs coming from the same region have similar functional implications during transcriptional regulation, we considered such circRNAs as common candidates between tumor and normal-adjacent tissues. Therefore, if a circRNA was identified in tumor and not in the normal-adjacent tissue, we termed such candidates as tumor-specific circRNAs and found 256, 288 and 411 tumor-specific circRNAs in TN, ER+ and HER2+ breast cancer subtypes respectively.

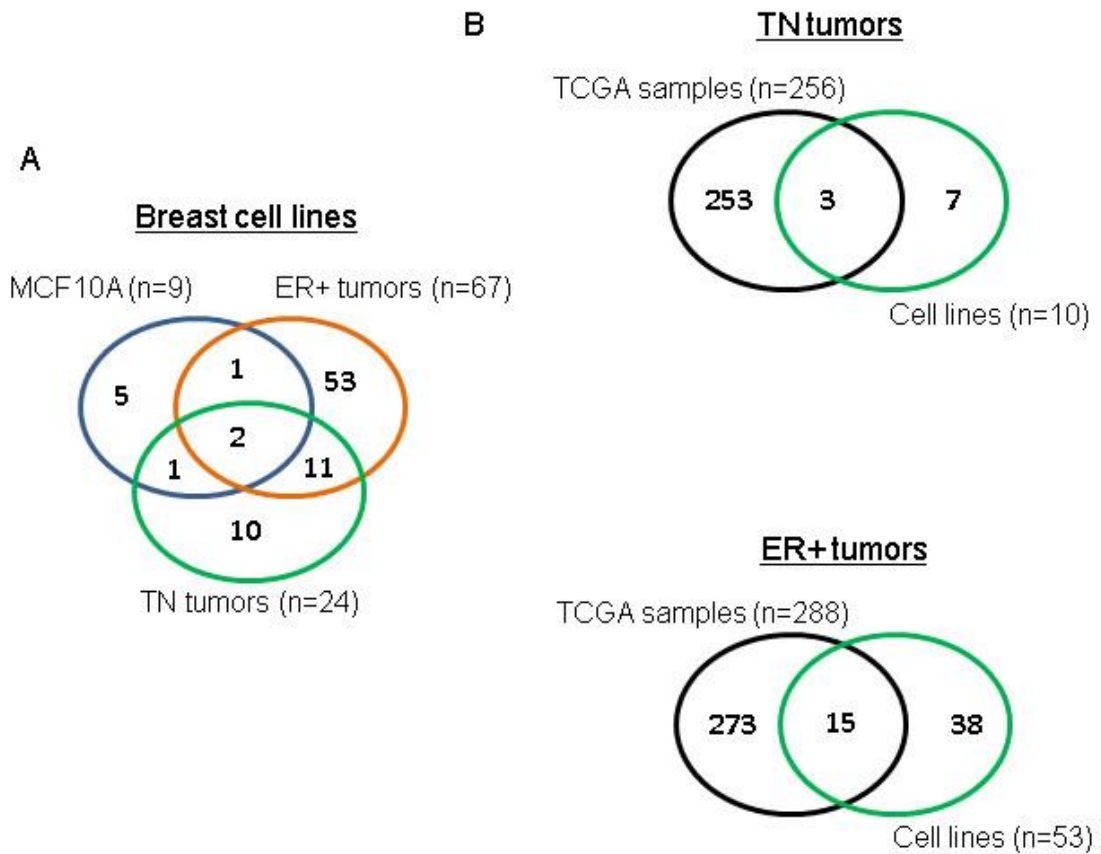
Because the number of normal-adjacent samples was much smaller than the number of breast tumor samples (most tumor samples did not have a paired

normal-adjacent tissue sample), we also calculated the ratio of unique circRNAs to the number of samples. Interestingly, after normalization, we see that circRNAs have a higher count in normal-adjacent samples, as shown in Table 6. We tested the significance of this observation using ANOVA and found that normal-adjacent samples of ER+ subtype had a p-value  $< 8.96e-06$  compared to the tumor. However, for TN and HER2+ subtypes the probability measure was insignificant (p-value  $> 0.05$ ), and combining all subtypes also did not show a significant increase in a number of normal-adjacent tissue circRNAs (p-value 0.11).

#### **3.2.4.1.2 *Tumor-specific circRNAs in breast cancer cell lines also present in breast cancer tissues***

circRNAs from the TN and ER+ cancer cell lines were compared to those from the non-tumor MCF10A breast cell line (see Table 5 for subtype classification of cell lines; no HER2+ cell lines were available). This comparison yielded 10 TN-specific and 53 ER+ -specific circRNAs (Figure 11A). We checked for common tumor-specific circRNAs between breast cancer cell lines and breast cancer TCGA samples. We also compared these circRNAs to the 256 circRNAs identified earlier in TCGA TN breast cancer samples and the 288 circRNAs obtained from TCGA ER+ breast cancer samples. As shown in Figure 11B, we found that three circRNAs were shared between TN breast cancer cell lines and

TCGA TN breast cancer samples, and 15 circRNAs were shared between ER+ breast cancer cell lines and TCGA ER+ breast cancer samples.



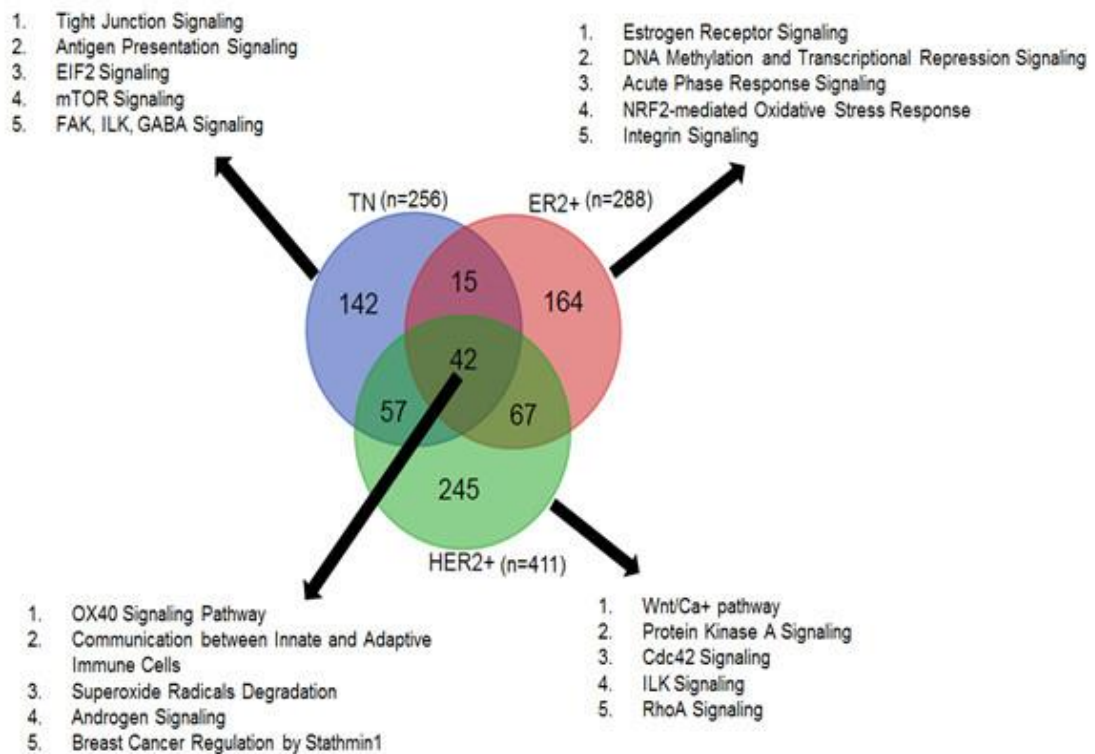
**Figure 11:** TCGA tumor-specific circRNAs also found in breast cell lines. (A) overlap of circRNAs between different subtypes for breast cell lines, (B) overlap of TN and ER+ tumor-specific circRNAs between TCGA and cell lines.

### **3.2.4.1.3 Tumor-specific circRNAs are associated with cancer-related canonical pathways**

The TN, ER+, and HER2+ breast cancer subtypes have unique prognostic and therapeutic characteristics. Although the gene expression profiles of these subtypes are markedly different [95, 96], a shared population of genes behaves similarly across them. We observed a comparable trend for circRNAs. We found that 42 tumor-specific circRNAs were common across TN, ER+, and HER2+ subtypes. At the same time, we also observed 142 TN, 164 ER+ and 245 HER2+ tumor-specific circRNAs that are exclusive to each subtype.

Because circRNAs have post-transcriptional regulatory functions and tend to influence overlapping or neighboring genes [16, 77], we annotated the tumor-specific circRNAs with protein-coding genes using the Ensembl reference system (version GRCh37.75). Pathway analysis demonstrated that most tumor-specific circRNAs were associated with cancer-related canonical pathways. The 42 circRNAs common to all three breast cancer subtypes were annotated with 45 genes, of which 33 genes ( $p$ -value =  $8.43E-05$ – $4.09E-03$ ) were associated with cancer-related pathways. As shown in Figure 12, these circRNAs are likely involved in various hormone signaling, immune cell communication, and OX40 signaling pathways. The circRNAs ( $n = 142$ ) unique to TN tumor samples were linked to a total of 370 genes of which 220 genes ( $p$ -value =  $7.79E-06$ – $1.26E-02$ ) were associated with cancer pathways such as tight junction, antigen presentation, and mTOR signaling pathways. Likewise, HER2+-specific circRNAs

(n = 245), annotated with over 1,500 protein-coding genes, had 855 cancer-related genes (p-value = 1.65E-14–2.24E-03) involved in Wnt signaling, Cdc42, and ILK signaling pathways. The ER<sup>+</sup>-specific circRNAs (n = 164) were found to overlap and/or neighbor 170 genes of which 129 cancer-related genes (p-value = 2.28E-12–6.82E-03) were associated with estrogen receptor signaling, epigenetic signaling, and oxidative stress response pathways.



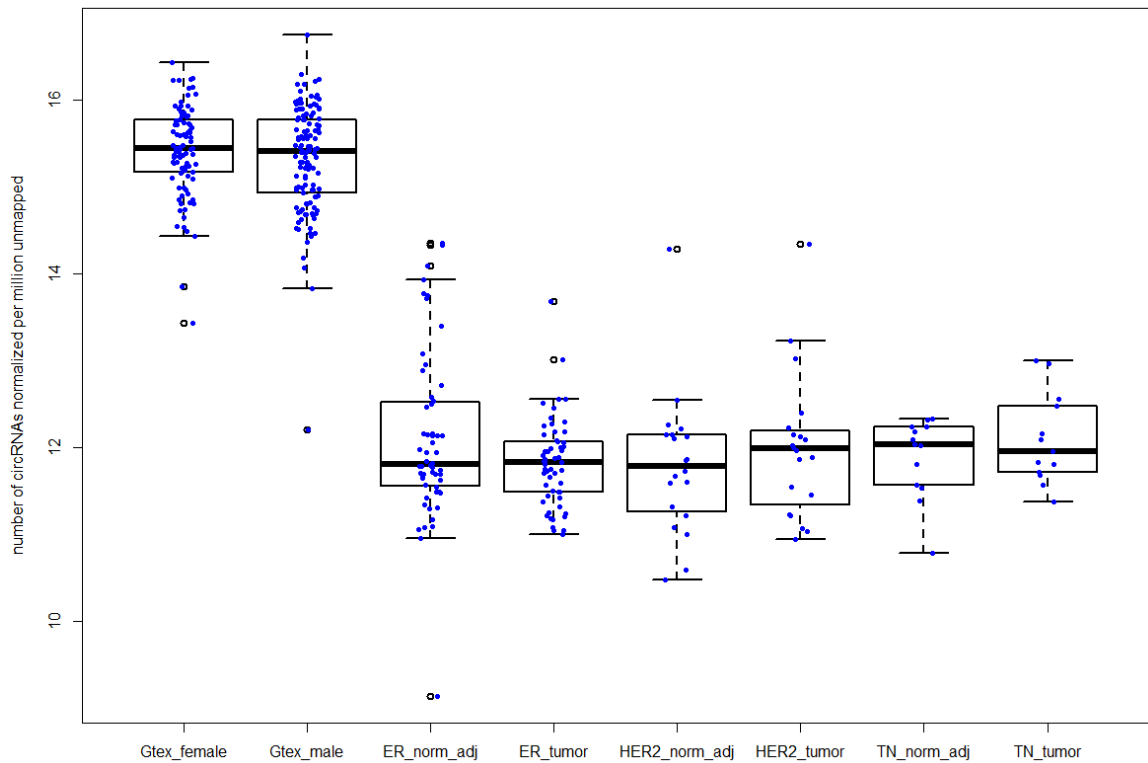
**Figure 12:** Tumor-specific circRNAs common and unique to TN, ER<sup>+</sup> and HER2<sup>+</sup> subtypes and the top canonical pathways associated with each subtype.

### **3.2.4.2 Paired normal-adjacent tissue analysis**

#### **3.2.4.2.1 *Normal-adjacent samples have more unique circRNAs than tumor samples in ER+ subtype***

We obtained paired breast tumor and normal-adjacent data from TCGA for 13 TN, 56 ER+, and 20 HER2+ samples. The circRNA results showed that the normal-adjacent samples had a higher number of unique circRNAs than the matched tumors in 5/13 TN patients, 23/56 ER+, and 6/20 HER2+ samples. Using standard paired t-test, again we found that in ER+ cancer, number of circRNAs was higher in normal-adjacent than tumor with p-value < 0.027. No correlation was observed between number of unmapped reads and circRNA number ( $R^2 = 0.099$ ), and after normalizing for unmapped reads, we still observed significant difference (p-value < 0.041) between ER+ normal-adjacent tissue and tumor samples. The TN and HER2+ patients did not show significance, p-value > 0.05 and combining all subtypes (89 pairs) yielded p-value < 0.1.

A large number of circRNAs were observed in normal breast samples from GTEx data. To confirm that number of circRNAs observed in normal samples is higher than breast tumors, we analyzed an independent cohort of 218 normal breast mammary tissues from the GTEx project (<http://www.gtexportal.org/home/>). After normalizing for library size, we observed a higher number of circRNAs compared to all three TCGA breast subtypes (Figure 13).

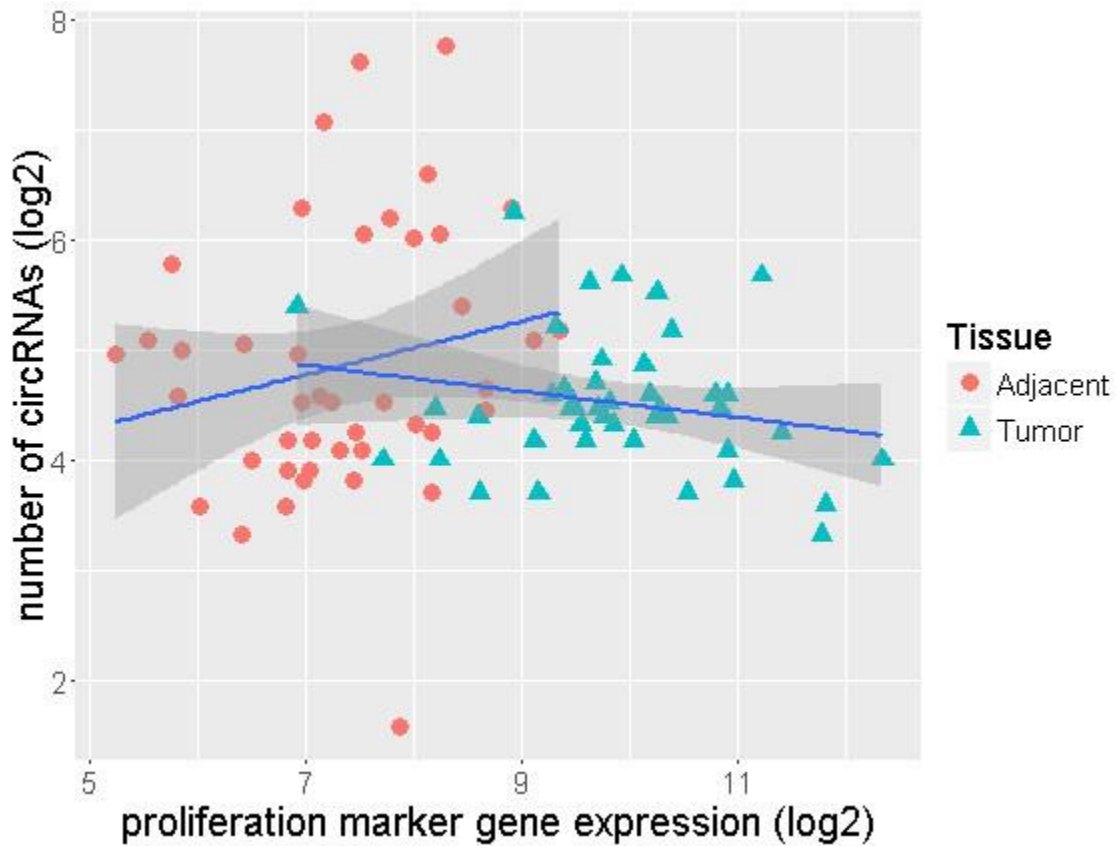


**Figure 13:** Increased number of circRNAs in normal breast samples compared to breast tumor subtypes in TCGA. A legend from left to right – Gtex\_female and Gtex\_male represent female and male mammary tissues from the Gtex project; ER+, HER2+ and TN normal-adjacent and matched tumors from TCGA are represented by ER\_norm\_adj, ER\_tumor, HER2\_norm\_adj, HER2\_tumor, TN\_norm\_adj and TN\_tumor respectively.

Recently, Bachmayr-Heyda et al. [91] reported that total number of circRNAs is negatively correlated with tumor proliferation marker MKI67 in colorectal cancer. Here we used a collection of 11 genes: BIRC5, CCNB1, CDC20, CEP55, MKI67, NDC80, NUF2, PTTG1, RRM2, TYMS, and UBE2C that are signatures for proliferation and are also part of the PAM50 classification gene panel [97]. We calculated the risk-of-relapse proliferation score (ROR-P) [93] for these genes to see if they have similar negative correlations with breast cancer subtypes.

We observed that ER+ normal-adjacent tissue samples had a higher number of circRNAs and displayed lower levels of proliferation marker gene expression than ER+ tumor samples. Figure 14 is plotted between the ROR-P score and circRNA numbers for the tumor samples and indicates that the number of circRNAs in the ER+ tumors tends to decrease with average increase in gene proliferation. This trend is explained by a slightly negative correlation of  $-0.22$ . A corresponding analysis of paired HER2+ and TN samples revealed correlations of  $-0.15$  in HER2+ and  $0.24$  in TN tumors.



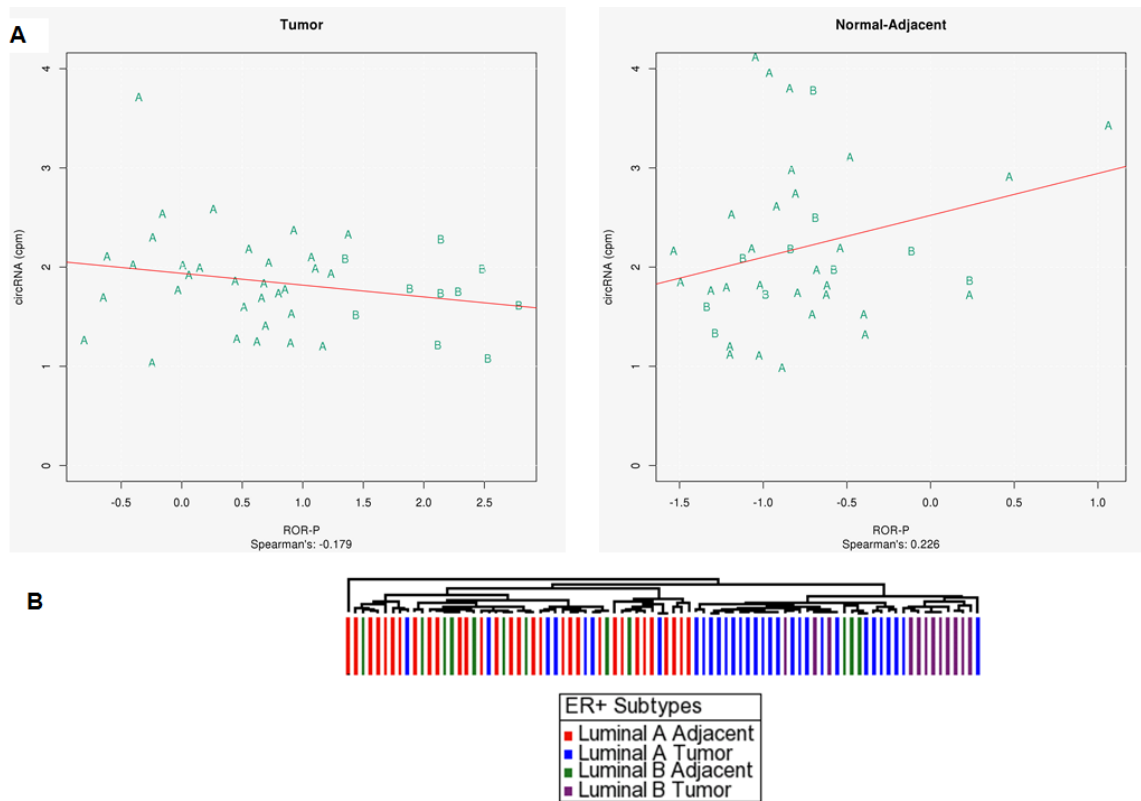


**Figure 14:** Lower number of circRNAs as gene proliferation increases in ER+ tumor samples

**3.2.4.2.2 ER+ luminal A and luminal B tumor tissues have distinct proliferation patterns based on the number of circRNAs**

Because circRNAs appear to be promising markers for proliferation in ER+ tumors, and since a number of circRNAs were significantly different between normal-adjacent and tumors, ( $p$ -value < 0.027), we further investigated if they could distinguish between the luminal A and luminal B types, as luminal B tumors proliferate more rapidly. First, we used PAM50 centroid modeling to identify the

tumor and normal-adjacent Luminal A and B subtypes for TCGA ER+ samples. Next, using all matched tumor and normal-adjacent ER+ samples (56 pairs), we plotted the number of circRNAs with respect to tumor proliferation. A clear distinction between the two ER+ types was evident for the tumor samples (Figure 15A). Luminal B tumors had fewer circRNAs (18 on average) than Luminal A tumors (25 on average) and this difference in circRNAs number was significant with  $p$ -value  $< 0.011$  using Welch t-test. Luminal B normal-adjacent samples had a similar number of circRNAs to luminal A normal-adjacent samples –24 and 30 on average, respectively, which was not statistically significant ( $p$ -value = 0.31). An unsupervised hierarchical clustering analysis, shown in Figure 15B, also indicated that tumor and normal-adjacent samples cluster separately based on their circRNA numbers. In addition, Luminal B tumors separated out into their own sub-cluster within the tumor arm. These results suggest that Luminal A and B tumor samples show distinct differences in terms of proliferation marker gene expression based on their circRNA numbers. We hypothesize that this measure may be of use for other cancers with heterogeneous subtypes.



**Figure 15:** (A) Luminal A and Luminal b tumor samples show distinct separation based on their circRNA numbers when plotted against tumor proliferation, (B) Unsupervised hierarchical clustering analysis shows separation of Luminal A and Luminal B tumor and adjacent samples based on their circRNA numbers.

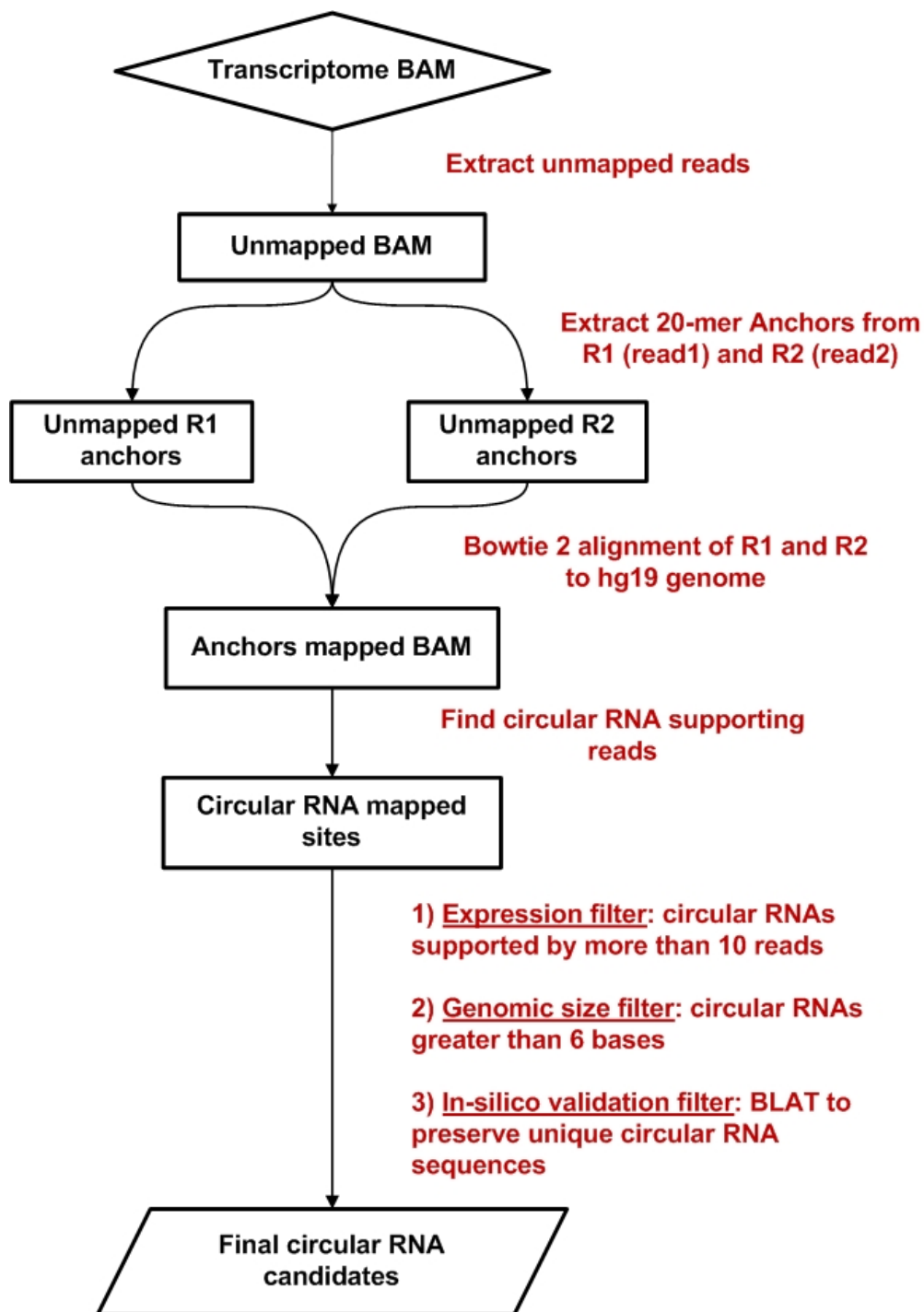
### 3.3 Methods

#### 3.3.1 Circ-Seq workflow

The Circ-Seq workflow flowchart is represented in Figure 16. Circ-Seq is an extension of the circRNA detection methodology by Memczak et al. [16] and incorporates essential filters as well as a comprehensive annotation to the final list of circRNA candidates. Circ-Seq starts by fragmenting unmapped reads from the aligned transcriptome BAM file into short 20-mer anchors from their 5' and 3'

ends and are then realigned against the reference genome. For every unmapped read, if the anchor pair maps in a 3' to 5' fashion, the alignment is shortlisted as possible evidence for a circRNA. Next, acceptor and donor splice sites, i.e., AG and GT, are checked for the selected 3' and 5' anchors. The presence of anchors within the splice sites is treated as initial confirmation of the fusion of exons in a circRNA fashion. At this point, the workflow quantifies the number of anchors supporting each circRNA candidate.

Next, three unique filters are applied to eliminate unexpressed and false-positive circRNAs: an expression filter, a genomic size filter, and a validation filter. The expression filter retains circRNA candidates supported by a sufficient number of junction-spanning reads and is set to 5 reads by default, which is considerably more stringent than existing approaches [16]. The genomic size filter discards any candidates shorter than six bases. Finally, to ensure that circRNAs reported by the workflow are not identified from repetitive regions of the genome, the validation filter uses BLAT [51] to confirm that the 3' (head) and 5' (tail) coordinates of the circRNA represent unique locations of the genome. After completing the analysis, the workflow provides a circRNA quantification report and a FASTA file that contains 50-base nucleotide sequences containing the 3'–5' fused junction of all circRNAs identified.



**Figure 16:** Circ-Seq bioinformatics workflow flowchart.

### **3.3.2 TCGA breast cancer transcriptome data**

We downloaded 1,034 breast cancer RNA-Seq binary alignment map (BAM) files from the TCGA Research Network (<http://cancergenome.nih.gov/>) using the National Cancer Institute (NCI) Genomic Data Commons (GDC) resource (<https://gdc.cancer.gov/>). The un-stranded Illumina TrueSeq protocol was used to obtain 50 nucleotide paired-end reads from TCGA breast cancer RNA-Seq samples. The paired-end reads were then aligned using MapSplice v12\_07 [43]; these reads contained both reads mapped to the human reference genome (hg19 / NCBI 37.1) and unmapped reads.

### **3.3.3 TCGA breast tumor and normal-adjacent samples and normal breast mammary tissue from GTEx**

We obtained clinical metadata for the 1,034 breast cancer samples from the NCI GDC Data Portal (<https://gdc-portal.nci.nih.gov/>). Because TCGA continues to add breast cancer cases to its repository, the most recent number of breast cancer samples available from TCGA may be higher than the number used in this work. We first classified the samples into the three predominant molecular subtypes – TN, ER+, and HER2. Out of 1,034 samples, we were able to classify subtypes for 885 samples of which 561 were ER+, 141 were TN, and 183 were HER2+ samples.

We downloaded BAM files for 218 normal breast samples (126 male and 92 female samples) from the Gtex project (<http://www.gtexportal.org/home/>) using Aspera client (<http://asperasoft.com/>). Samples were sequenced using Illumina TrueSeq paired-end RNA sequencing with read length 75 bp. The transcriptome BAM files that were downloaded for the 218 samples were aligned to the hg19 reference genome using Tophat [59].

### **3.3.4 Breast cancer cell lines**

We also obtained RNA-Seq paired-end sequence files for six breast cancer cell lines (BT20, BT474, MCF7, MDAMB468, T47D, and ZR751) and one cell line derived from normal breast cells (MCF10A) [94]. Sequences from the cell lines were processed using the Mayo Analysis Pipeline for RNA Sequencing (MAP-RSeq) to yield BAM files for use with the Circ-Seq workflow [37]. The number of unmapped reads for the cell lines varied from 5 to 22 million reads.

### **3.3.5 Pathway analysis for tumor-specific circRNAs**

Gene names and annotations of genes that either overlap or neighbor tumor-specific circRNAs were obtained using the Ensembl reference system (version GRCh37.75). Enriched canonical pathway analysis for tumor-specific circRNAs in the breast molecular subtypes was performed using the Ingenuity pathway analysis software IPA (Ingenuity® Systems, <http://www.ingenuity.com>). Biological functions and diseases information within the IPA software was used for critical

investigation of cancer-related pathways. Open source analysis toolkit WebGestalt [98] was also used to derive pathway results.

### **3.3.6 CircRNA validation**

MCF7 breast cancer cells (American Type Culture Collection Manassas, VA) were cultured in EMEM medium containing 10% fetal bovine serum (FBS) at 37°C in 5% CO<sub>2</sub>. Total RNA and genomic DNA were isolated using the RNeasy Plus Micro Kit and DNeasy Blood & Tissue Kit (QIAGEN, Inc., Valencia, CA) respectively. DNA and RNA quality was analyzed using the NanoDrop 8000 spectrophotometer. qRT-PCR was performed with the Power SYBR® Green RNA-to-CTTM 1-Step Kit (AB, Foster, CA) using a Stratagene Mx3005P Real-Time PCR detection system. GAPDH DNA and RNA were used as controls for the experiment. We designed two sets of primers, convergent primers that bound to linear 5'–3' mRNA transcripts and divergent primers that bound to the circRNA transcript (chr14:102,466,325–102,500,789) formed in a 3'–5' fashion, which were provided by Integrated DNA Technologies. After gel purification using the QIAquick Gel Extraction Kit (QIAGEN), the qRT-PCR product was sequenced using the Sanger method to confirm the head-to-tail splicing.



### 3.4 Discussion

In this study, using existing bioinformatics approaches defined by Memczak et al. [16] we developed a comprehensive analytical workflow called Circ-Seq. We also introduced three essential filters for identification and characterization of stable and expressed circRNAs in Circ-Seq. The workflow was designed with flexibility to allow users to configure these filters based upon their choice to report results that are either stringent or lenient. Circ-Seq is also designed with speed in mind. It is built to work on a multi-threaded cluster environment and can analyze numerous samples in parallel at any given time.

Circ-Seq was applied to the transcriptome of 885 TCGA breast cancer samples, and we identified numerous circRNAs unique to breast tumors and normal-adjacent tissues. To our knowledge, this is the first report to catalog circRNAs unique to the TN, HER2+, and ER+ molecular subtypes of breast cancer, as well as circRNAs common to all of the subtypes but absent from normal-adjacent tissue. Finally, using a panel of 11 tumor proliferation marker genes in combination with circRNA abundance, we show that circRNA number is associated with tumor proliferation and that luminal A and luminal B tumors have distinct representations of circRNA numbers within the ER+ breast molecular subtype.

We also identified circRNAs in the breast cell lines and were able to successfully validate the largest circRNA identified in MCF7 found at genomic location chr14:102,466,325–102,500,789 with 12 supporting junction reads. This circRNA was a spliced product of gene DYNC1H1 and spanned from exons 17 to 56 of the gene. Considering that the exon-intron structure remains intact, the size of this circRNA is about 7 kb and may play a role in post-transcriptional regulation. Notably, the circRNA contained microRNA response elements (MRE) for miR-150 and miR-661 with 29 and 23 unique binding sites respectively. These two microRNAs have been previously reported to have associations with cancer [99, 100]. In searching for other microRNAs that have over 20 binding sites, we found non-conserved microRNAs such as miR-3613, miR-4731 and miR-5095, each contains 25 MRE sites along the circRNA. It is possible that since the circRNA contains several binding sites for microRNAs, this could be a candidate player in breast cancer competing endogenous RNA (ceRNA) networks.

Recent studies suggest that circRNAs have other functions that are more common than the microRNA sponge effect. Notably, circRNAs are shown to participate actively in pre-mRNA splicing [80] and also as active promoters of transcription of parent mRNAs [86]. We believe that the circRNAs reported in this study can also have implications similar to the above functions in breast cancer.

Although validation results suggest that the workflow reported legitimate circRNAs, the reliability of the workflow and the measure of the false positive rate can only be determined based on its application to more transcriptome datasets and validation of results in future. The number of unmapped reads is a key player in identifying circRNAs within a sample. We observed that unmapped reads for the breast tumor and non-tumor cell lines range was 5–22 million and the range for TCGA samples was 5–78 million. Samples with unmapped reads at the low end of the spectrum can likely have a correspondingly low number of circRNAs reported. We hypothesize that the number of circRNAs identified for BT20, T47D, MCF7, and MCF10A was artificially low due to the small number of unmapped reads available for these samples.

One of the limitations of this study is that the RNA-Seq libraries from TCGA are prepared using Illumina TruSeq, which enriches for poly-A tail transcripts [101], thus substantially limiting the number of circRNAs detected. Despite this limitation, we identified large numbers of circRNAs in the TCGA breast cancer data. Stranded total RNA and RiboMinus libraries may improve the detection of circRNAs [16, 72, 73, 82, 84, 87]. We acknowledge that the circRNAs identified here are only a small subset of the actual population of the circRNAs present in breast cancer samples. Because the number of circRNAs detected increases with the number of samples investigated, as shown in Table 6, the number of circRNAs detected for the TN, and HER2+ subtypes are probably

underestimated due to their smaller sample size. This could also be indicative of why we observed poor correlations and non-significant probability measures for these subtypes when the corresponding associations always held true for ER+ samples. Likewise, it is uncertain at this point whether the tumor proliferation analysis for TN and HER2+ patients with matched tumor and normal-adjacent tissues would indeed have a negative correlation with circRNA numbers or not, if an adequate number of samples were available for these subtypes.

Taking together the biological complexities of cancer, individual RNA classes cannot be considered in isolation. Cooperative communication between different types of noncoding RNAs and protein-coding genes or mRNAs exists [23, 25, 102, 103] which eventually tune the expression of target genes. In cancer, regulated expression of tumor suppressors and oncogenes is critical to tumorigenesis. Competing endogenous RNA networks comprising of complex interactions between mRNA, microRNA and circRNA molecules can greatly influence the post-transcriptional activity of such genes. mRNA stability, or lack of stability—depending on how the circRNAs and microRNAs interact via microRNA binding sites—can significantly impact gene expression, with serious repercussions for tumorigenesis. Innovative and ingenious bioinformatics techniques need to be developed that can unravel ceRNA crosstalk between such RNA types and eventually lead to novel findings which can be used as potential diagnostic targets to improve treatment of cancer. It is possible that the

findings that emerge from the study of circRNAs will lead to improvements in the diagnosis and treatment of complex, heterogeneous diseases such as cancer.

# **Chapter 4: ReMix - A novel bioinformatics approach to integrate mRNA-microRNA interactions using MRE frequencies from RNA-Seq data**

## **4.1 Introduction**

The transcriptome activity in a tumor or normal cell involves complex interactions between mRNAs (mRNAs) and other coding and non-coding RNAs that greatly influence the post-transcriptional availability of genes in the cell. One such interaction, which has shown to have enormous implications in cancer, is the interaction between mRNAs and microRNAs [26, 28, 104]. MicroRNAs regulate gene expression by using their seed sequences (6 - 8 bases long) to bind to the microRNA response element (MRE) sites located on the 3' untranslated regions (3' UTR) of mRNAs. Genes can have one or more distinct MRE sites, thus being targets to multiple microRNAs, and likewise, microRNAs can bind to MRE sites of several different gene targets [105, 106]. Most of the times microRNAs bind partially, and sometimes entirely, to MRE sites of their target leading to suppressed translation or mRNA degradation. Lately, it has been shown that non-coding circRNAs (circRNAs) also contain conserved binding sites for

microRNAs and thus can function as critical post-transcriptional regulators [16]. CircRNAs can use their MRE binding sites to compete with mRNAs for a shared pool of microRNAs. For heterogeneous diseases such as cancer, it is critical to understand how complex interactions between mRNAs, circRNAs, and microRNAs impact the levels of target mRNA expression in the tumor. The mRNAs belong to diverse biological pathways and thus alterations in target gene expression via microRNA binding as well as regulation via circRNA competition can impact several cellular processes, such as cell cycle, proliferation, cell death, apoptosis, etc., during cancer development, progression and migration. Thus, finding key players among the mRNA-microRNA-circRNA interacting networks can yield identification of new biomarkers in cancer, especially for cancer subtypes that are least responsive to current modalities of treatment.

Eminent advancement in NGS technologies has allowed consortia such as The Cancer Genome Atlas (TCGA) to sequence large populations of cancer patients and to make data available for research. Studies have used microRNA and mRNA expression profiles across many cancer types in TCGA to infer microRNA-target interactions that could be active and functional in different cancer types [107]. Other studies have looked into alternative polyadenylation of 3'UTRs in TCGA bladder cancer to show how short 3'UTR lengths affect mRNA stability and attenuate protein translation [108]. There have also been focused

analyses of SNPs present at 3'UTR regions of genes that affect microRNA binding and are shown to be associated with multiple cancer subtypes [109]. However, no work has been done so far to extrapolate TCGA RNA-Seq data alone to analyze MRE sites across the transcriptome of the tumor and normal samples to obtain insights into unique interactions between mRNAs and microRNAs at the 3' UTR regions in the tumor.

Here we present a new bioinformatics approach called **ReMix** (pronounced “remix”) – mRNA-MicroRNA Integration, that leverages RNA-Seq data to quantify microRNA binding sites (MREs) at the 3' UTR of genes across the transcriptome. ReMix quantitates MRE sites in tumor and normal separately, which later enables identification of differentially expressed MREs that are statistically significant in the tumor. Since MRE is the link between mRNAs and microRNAs, ReMix effectively reports candidate mRNAs and microRNAs that have a differential pattern of binding in the tumor. Additionally, based on our previous work in circRNAs [38], we also integrated circRNA transcripts for mRNAs with differential tumor MRE sites. Finally, bringing together a) mRNAs with tumor-specific MRE sites, b) microRNAs that have the potential to bind to these MRE sites, and c) circRNA transcripts for these targets, ReMix reports potential mRNA-miRNA-circRNA candidates that have unique interactions specific to the tumor. The method can be applied to study any cancer subtype of complex disease with tumor/affected and normal sequencing datasets. As an example,



here we have shown the application of ReMlx method to TCGA RNA sequencing data [110, 111] for Triple Negative Breast Cancer (TNBC) cases, normal-adjacent tissues, along with ER+ and HER2+ tumors, to identify ReMlx candidates unique to the TNBC tumors. Our analysis of the TCGA breast cancer cohort has identified promising targets for better diagnosis and treatment of TNBC disease.

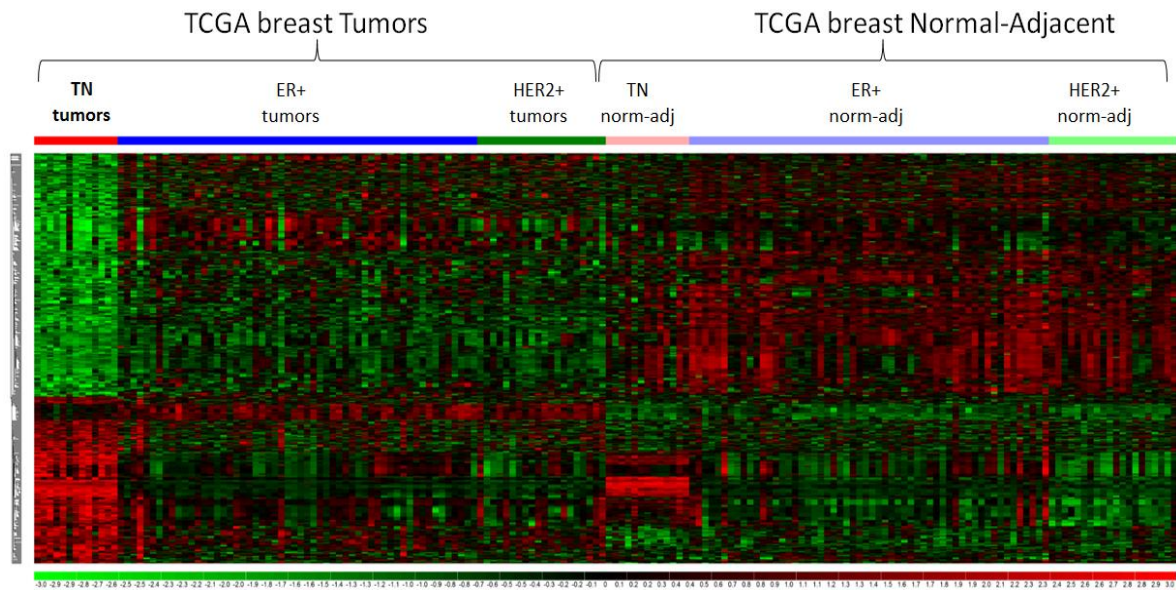
## **4.2 Results**

### **4.2.1 Application of ReMlx to triple-negative tumor and normal-adjacent pairs**

The RNA-Seq aligned BAM files for 13 pairs (Tumor and Normal-Adjacent) cases from TN subtype, 56 pairs of ER+ subtype and 20 pairs of HER2+ subtype were obtained from the MAPR-Seq workflow [37]. These BAM files were further subset to 3' UTR regions of individual genes (n=12,455 as per TargetScan v7.0) and converted to fastq format for input to the ReMlx workflow (see Methods). The pre-computed MRE sequences (n=329 microRNA seeds, as per TargetScan 7.0) were also provided as input to FIMO [112] to report a read count for individual MREs located on every gene. The MRE counts were then normalized by factoring the library size, 3'UTR lengths and 3'UTR GC content of individual genes. Since genes can contain multiple MRE sites, at the end of the MRE

quantification process, we obtained normalized counts for a total of 111,522 MRE sites in tumor and normal-adjacent samples separately, for each subtype.

Next, we sought to identify MRE sites that had unique and significant levels of expression (high or low) in the TN tumor cases when compared to ER+ tumors, HER2+ tumors as well as TN, ER+ and HER2+ normal-adjacent cases. Using Dunnett-Tukey-Kramer pairwise multiple comparison statistical test on these tumor and normal-adjacent cases across all subtypes (total of 6 groups), we found 614 TN tumor-specific MRE sites that were significant at  $p$ -value  $< 0.05$ . As shown in Figure 17, we observed that more than half of the 614 MRE sites had lower expression in TN tumors when compared to other subtypes as well as TN normal-adjacent cases.

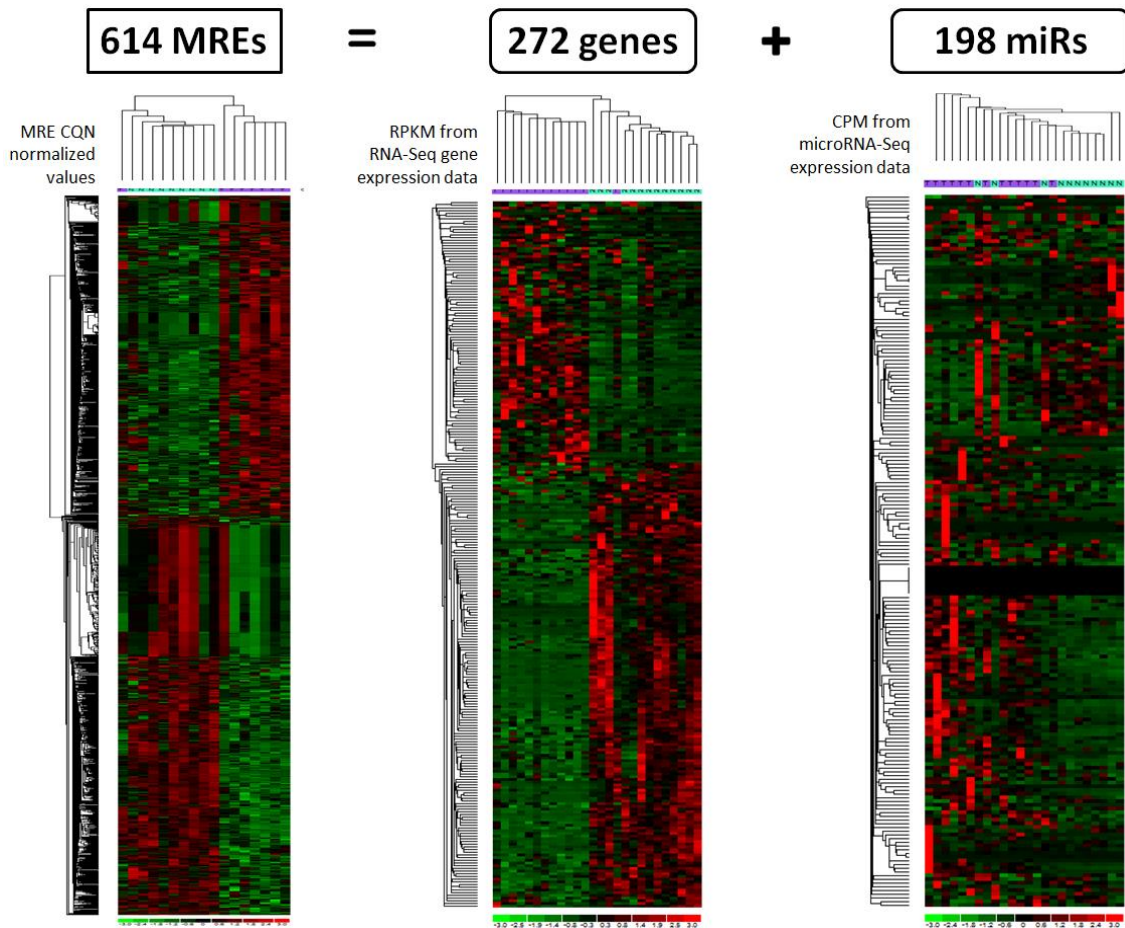


**Figure 17:** Heatmap of 614 TN-tumor specific MREs. The normalized CQN values of 614 MREs were obtained for TN, ER+ and HER2+ tumors and normal-adjacent (norm-adj) cases. As shown in the heatmap, these MREs have a distinct expression in TN tumors in comparison to the other subtypes as well as TN normal-adjacent cases.

#### 4.2.2 614 MREs associated to 272 genes and 198 microRNAs

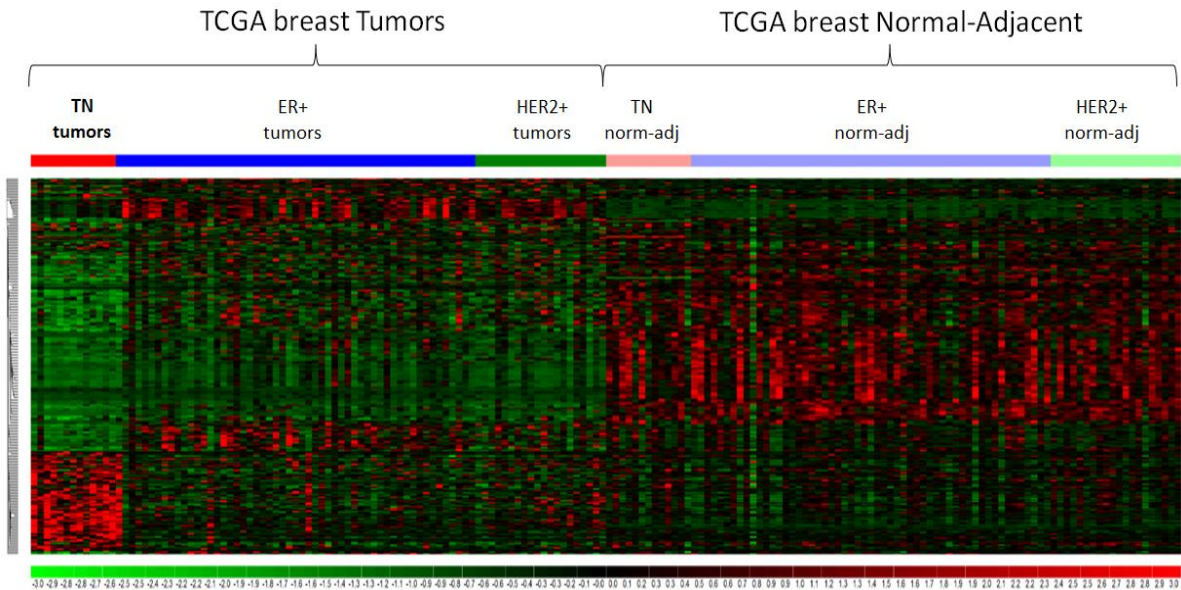
Since one MRE site is the combination of a gene and microRNA, the 614 TN tumor-specific MRE candidates were de-coupled to identify the respective genes and microRNAs. We obtained 272 genes and 198 microRNAs that were associated with these MREs. The heatmaps in Figure 18 represent the expression of the 614 MREs (CQN normalized values), 272 genes (RPKM values obtained from RNA-Seq data) and 198 microRNAs (CPM values obtained from microRNA-Seq data) for the 13 Triple-Negative Tumor and Normal-Adjacent samples. Un-supervised hierarchical clustering showed that the samples clustered well within their Tumor and Normal-Adjacent groups for all three data

types. Differential patterns of expression for the 614 MREs were also reflected in the heatmaps for the 272 genes and 198 microRNAs.



**Figure 18:** Un-supervised clustering and heatmap representation 614 TN-tumor specific MREs and their associated genes and microRNAs. The 614 MREs were associated with 272 genes and 198 microRNAs. CQN normalized values were obtained for the MRE sites. Likewise, RPKM normalized values from RNA-Seq and CPM normalized values from microRNA-Seq were obtained for these 13 pairs of TN tumor and normal-adjacent cases to generate their respective heatmaps. Unsupervised clustering of the cases indicates that tumor (purple) and normal-adjacent (cyan) were clustered well within their groups.

So far, based on MRE sites, we found that 272 genes showed evidence of differential expression at their MRE sites (614 MREs) for the TN tumors. Next, we were curious whether this MRE expression change for the 272 genes would continue to be reflective at the gene level and if they would still be distinct for the TN tumors when compared to other subtypes. For this we used RNA-Seq gene expression data for the TCGA cases (13 TN pairs, 56 ER+ pairs, and 20 HER2+ pairs) and found that, when looking across the ER+ and HER2+ subtypes, not all but a subset of the 272 genes still had high as well as low expression that was distinct to the TN tumors, as shown in Figure 19.

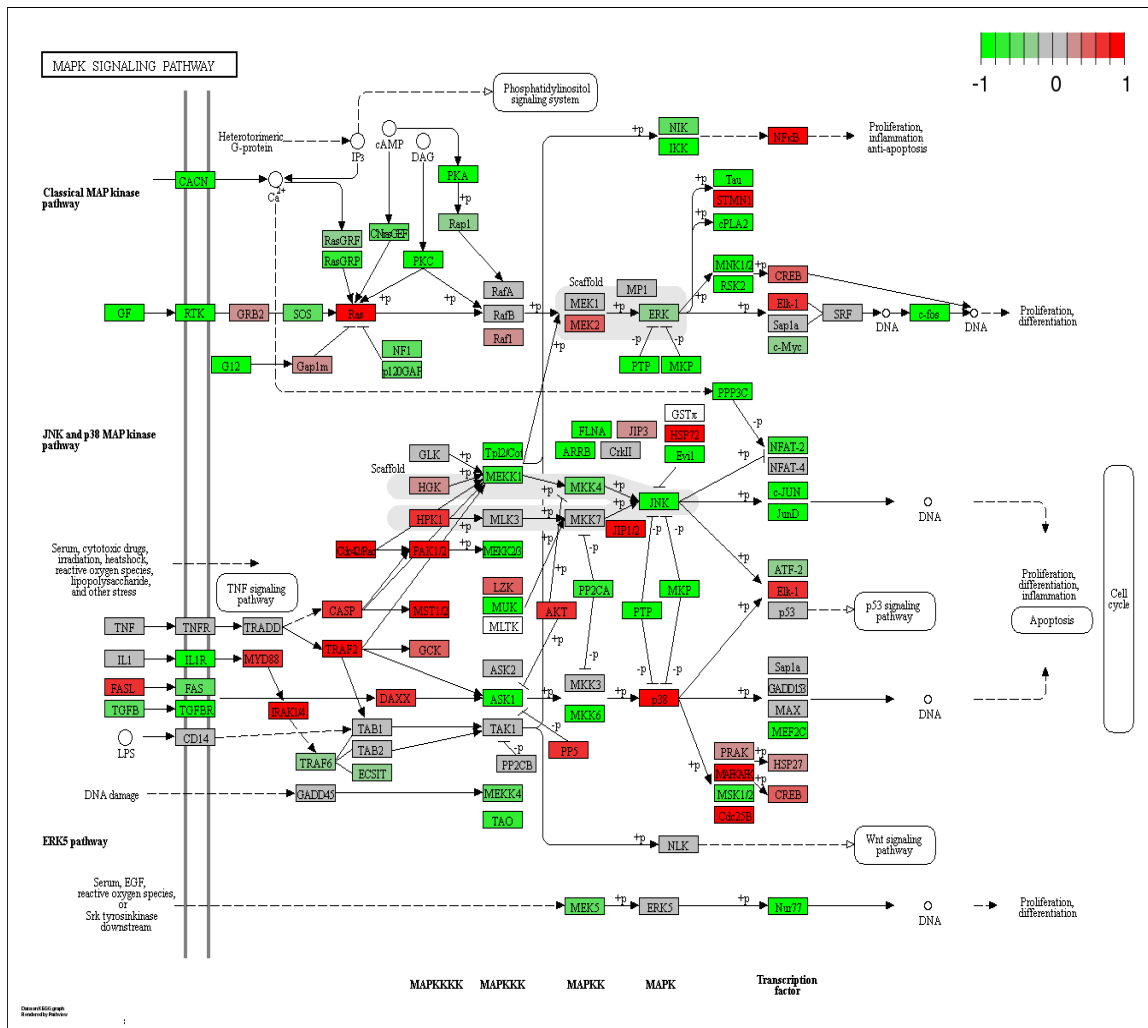


**Figure 19:** Heatmap of gene expression for 272 genes. Using supervised clustering, the RPKM normalized values were plotted for 272 genes from TN, ER+ and HER2+ subtypes. The heatmap indicates that a subset of these genes have different expression profiles in TN tumors when compared to other subtypes and normal-adjacent cases.

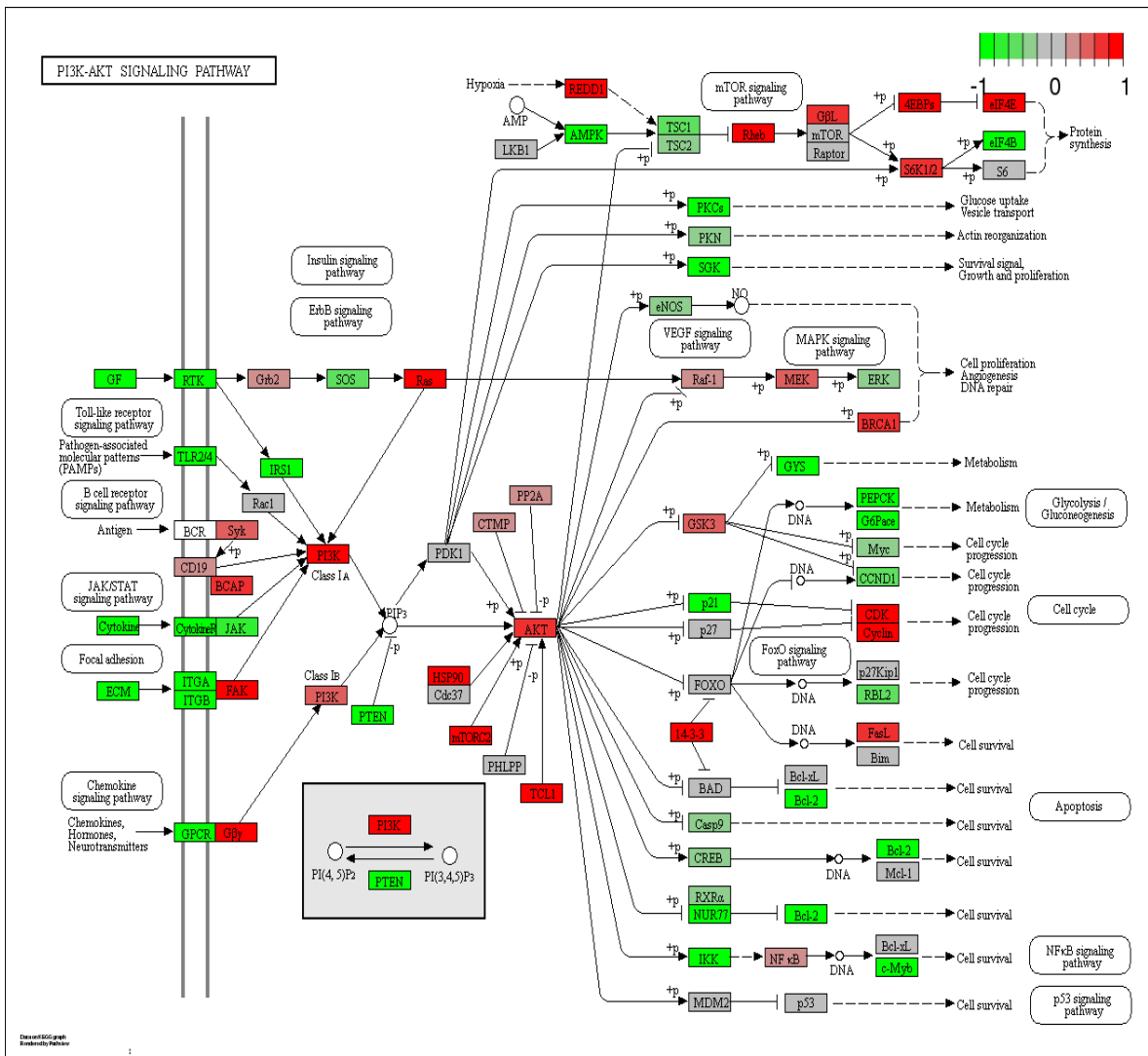
### **4.2.3 MAPK and PI3K-AKT signaling identified among top pathways**

The 272 genes obtained from MRE analysis were further analyzed to identify their associated canonical pathways. Using gene set enrichment analysis (GSEA) [113] on KEGG and REACTOME databases, we found the mitogen-activated protein kinase (MAPK) and phosphatidylinositol 3-kinase (PI3K-AKT) signaling cascades reported among the top significant pathways. Additionally, application of the Signaling Pathway Impact Analysis (SPIA) package also reported MAPK signaling pathway to be activated in these TN tumors.

Closer examination of the genes in the MAPK and PI3K-AKT pathways revealed that among several genes that were activated and repressed, oncogenes KRAS, NRAS, AKT, and NFkB were notably activated and tumor suppressor PTEN was repressed in the MAPK and PI3K-AKT pathways. Figures 20 and 21 illustrate the respective KEGG pathway diagrams for MAPK and PI3K-AKT signaling cascades.



**Figure 20:** MAPK signaling pathway. Genes in the pathway are colored based on their expression in TN tumors. Oncogenes NFKB and AKT are notably activated in this pathway.



**Figure 21:** PI3K-AKT signaling pathway. Genes in the pathway are colored based on their expression in TN tumors. Oncogenes PI3K, Ras, and AKT found activated and tumor suppressor PTEN repressed in this pathway.

MAPK and PI3K-AKT signaling pathways are huge network cascades with connections to several essential biological pathways downstream such as proliferation, cell cycle, glycolysis, apoptosis, protein synthesis, etc. As a result, these cascades comprise of large number of genes involved in conducting and



maintaining activities within the MAPK and PI3K-AKT pathways. From our MRE analysis, we know the gene-microRNA pairs that formed the 614 MRE sites that are distinct to TN tumors. We therefore investigated the relevant gene-microRNA pairs that were part of the MAPK and PI3K-AKT pathways. This information is tabulated in Table 7, where we found that 12 out of 294 genes (~ 4%) in the MAPK pathway and 13 out of 351 genes (~4%) in the PI3K-AKT pathway have MRE sites that have potential for differential binding of microRNAs in TN tumors. The microRNAs that bind to these differential MRE sites for respective genes are also provided in Table 7.

mRNAs	microRNAs
<b>MAPK signaling pathway</b>	
CACNA2D1	hsa-miR-429
PPP3CB	hsa-miR-330-5p; hsa-miR-486-5p
RASGRF1	hsa-miR-384
IGF1	hsa-miR-142-5p; hsa-miR-488-3p
HGF	hsa-miR-495-3p
EFNA5	hsa-miR-101-3p.2; hsa-miR-130b-3p; hsa-miR-489-3p; hsa-miR-96-5p
PDGFRA	hsa-miR-132-3p; hsa-miR-140-5p; hsa-miR-491-5p
FOS	hsa-miR-802
TGFBR2	hsa-miR-361-5p; hsa-miR-665
FLNC	hsa-miR-377-3p
ARRB1	hsa-miR-140-3p.1; hsa-miR-296-5p
PPM1A	hsa-miR-488-3p
<b>PI3K-AKT signaling pathway</b>	
DUSP1	hsa-miR-141-3p; hsa-miR-144-3p; hsa-miR-194-5p; hsa-miR-379-3p; hsa-miR-411-5p.1; hsa-miR-495-3p
IGF1	hsa-miR-142-5p; hsa-miR-488-3p
HGF	hsa-miR-495-3p
EFNA5	hsa-miR-101-3p.2; hsa-miR-130b-3p; hsa-miR-489-3p; hsa-miR-96-5p
PDGFRA	hsa-miR-132-3p; hsa-miR-140-5p; hsa-miR-491-5p
GHR	hsa-miR-129-5p; hsa-miR-132-3p; hsa-miR-505-3p.2
COL1A1	hsa-miR-129-5p; hsa-miR-133a-3p.2
THBS2	hsa-miR-182-5p; hsa-miR-379-3p
ITGA1	hsa-miR-27a-3p
PIK3R1	hsa-miR-126-3p.2; hsa-miR-361-5p; hsa-miR-455-3p.1; hsa-miR-488-3p;

	hsa-miR-493-5p
LPAR1	hsa-miR-142-3p.2
GNB4	hsa-miR-29a-3p; hsa-miR-381-3p
PTEN	hsa-miR-140-3p.2

**Table 7:** Gene-microRNA pairs with distinct TN-specific MRE sites that are part of the MAPK and PI3K-AKT pathways. This table lists genes that are a subset of the 272 genes obtained from the MRE analysis, which are members of the MAPK and PI3K-AKT signaling pathways. The microRNAs that bind to the MRE sites which were found to have distinct counts in TN tumors are also provided.

#### 4.2.5 CircRNAs associated with MAPK and PI3K-AKT pathways

Based on our previous work in TCGA breast cancer [38], we identified several circRNAs in TN breast tumors. The Circ-Seq workflow classified circRNAs as inter-gene and intra-gene circRNAs (based on genomic coordinates), and in this study, we focused on intra-gene circRNAs, which were annotated as circular transcripts formed from a single gene. We were curious if there was any evidence of circRNAs for the genes involved in the MAPK and PI3K-AKT pathways and found that indeed, a subset of genes also contained evidence of circular transcripts. Table 7 presents the candidate intra-gene circRNAs that were identified and the corresponding genes they are associated with, in the MAPK and PI3K-AKT signaling pathways.

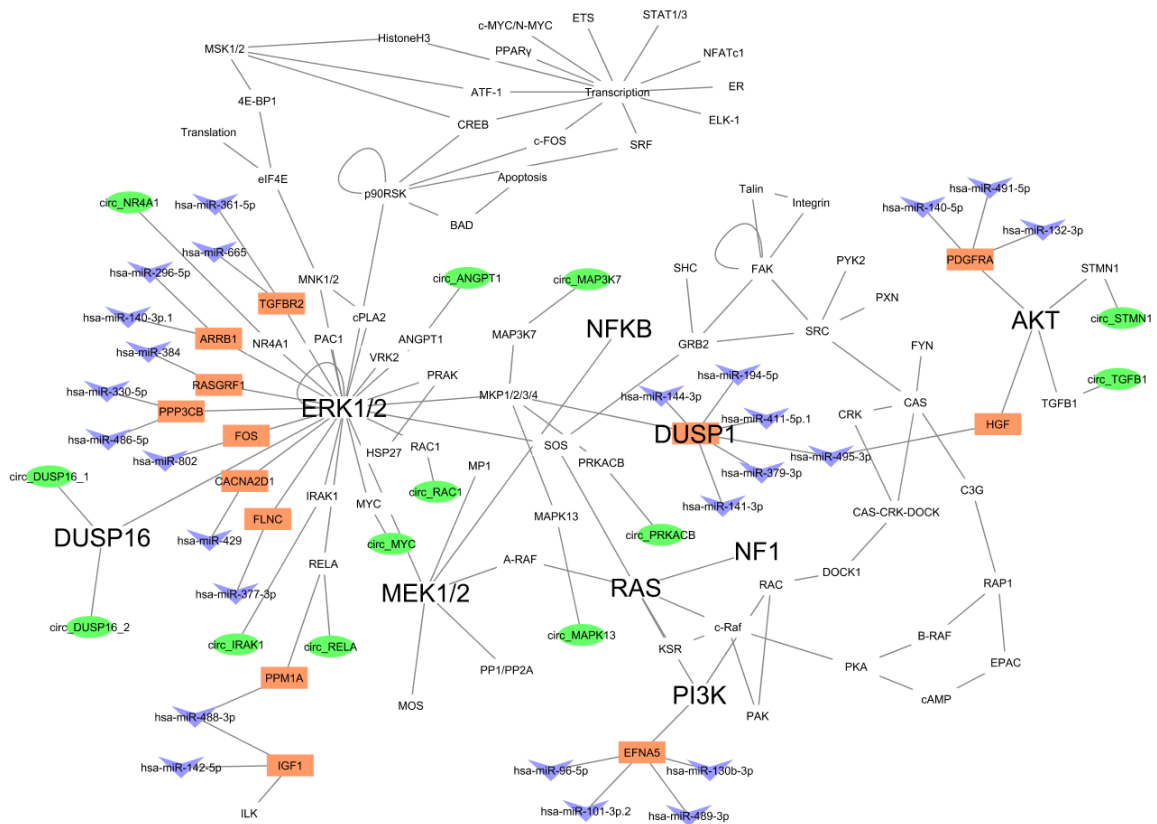
circRNA	Average expression in TN tumors	Gene (activated –A, repressed – R in TN tumors)	circRNA type
<b>MAPK signaling pathway</b>			
circ_PRKACB	3	PRKACB (R)	Tumor
circ_ANGPT1	5	ANGPT1 (R)	Tumor

circ_MYC	1.73	MYC (R)	Tumor
circ_STMN1	5	STMN1 (A)	Tumor
circ_TGFB1	1.75	TGFB1 (A)	Tumor
circ_RAC1	1.64	RAC1 (R)	Tumor-Specific
circ_IRAK1	1.53	IRAK1 (A)	Tumor-Specific
circ_MAP3K7	1.91	MAP3K7 (A)	Tumor
circ_MAPK13	3	MAPK13 (A)	Tumor
circ_DUSP16_1	2	DUSP16 (R)	Tumor
circ_DUSP16_2	3	DUSP16 (R)	Tumor
circ_NR4A1	1.5	NR4A1(R)	Tumor-Specific
circ_RELA	6	RELA (A)	Tumor
<b>PI3K-AKT signaling pathway</b>			
circ_ANGPT1	5	ANGPT1 (R)	Tumor
circ_RAC1	1.64	RAC1 (R)	Tumor-Specific
circ_FN1	1.61	FN1 (A)	Tumor
circ_SPP1	1.43	SPP1 (A)	Tumor
circ_TNC	1.67	TNC (R)	Tumor-Specific
circ_ITGB3	7	ITGB3 (R)	Tumor
circ_DDIT4	1.71	DDIT4 (A)	Tumor-Specific
circ_PPP2R1A	2.2	PPP2R1A (A)	Tumor
Circ_COL1A1	2.3	COL1A1 (A)	Tumor-Specific
circ_PPP2R5E	3	PPP2R5E (R)	Tumor
circ_HSP90B1	2.55	HSP90B1 (A)	Tumor
circ_CRTC2	1.44	CRTC2 (A)	Tumor
circ_GSK3B	12	GSK3B (A)	Tumor
circ_G6PC3	1.71	G6PC3 (R)	Tumor-Specific
circ_MYC	1.73	MYC (R)	Tumor
circ_CCND2	5	CCND2 (R)	Tumor
circ_MCL1	1.47	MCL1 (R)	Tumor
circ_NR4A1	1.5	NR4A1 (R)	Tumor-Specific
circ_RELA	6	RELA (A)	Tumor

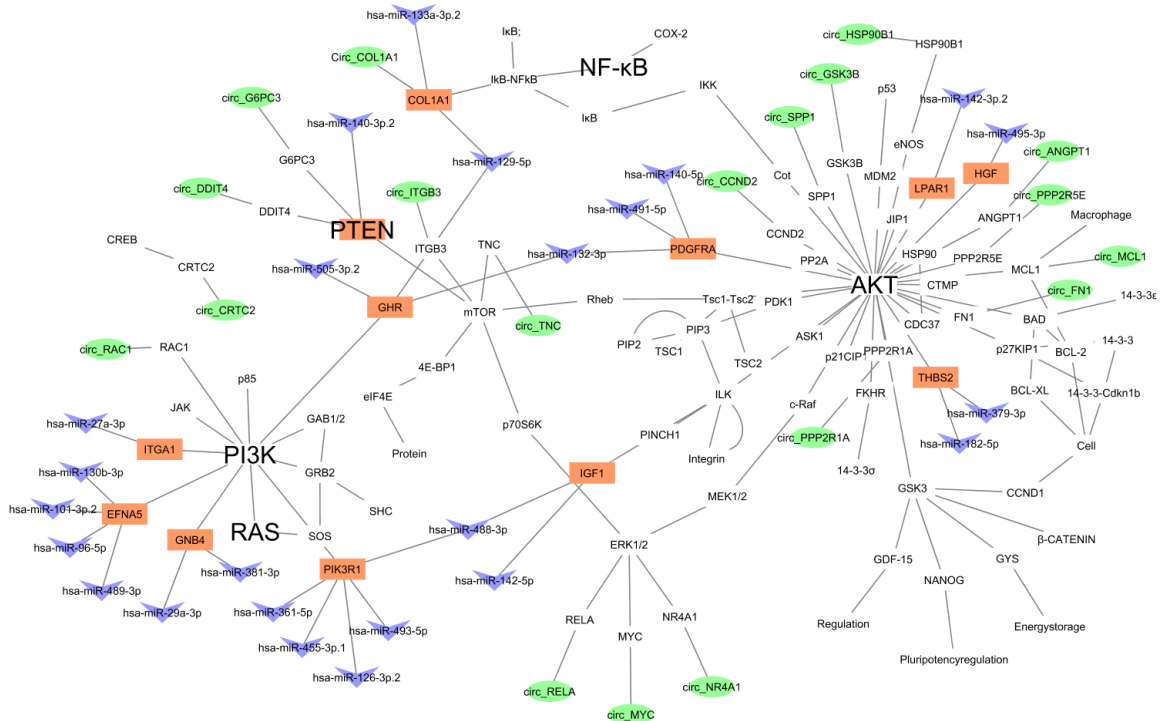
**Table 8:** Intra-gene circRNAs identified in MAPK and PI3K-AKT pathways. Circular transcripts obtained from the Circ-Seq workflow on TCGA TN breast cancer cases were used to identify intra-gene circRNAs associated with genes in the MAPK and PI3K-AKT pathways. CircRNA type ‘Tumor-Specific’ denotes that the circRNA was found only in TN Tumors, whereas circRNA type ‘Tumor’ indicates that there is a different circRNA isoform in the reported gene was found in TN normal-adjacent.

#### **4.2.6 mRNA-microRNA-circRNA interacting candidates in MAPK and PI3K-AKT pathways for TN breast cancer**

Based on the mRNA-microRNA results from the ReMix methodology as well as the circRNAs from Circ-Seq workflow, we expanded the MAPK and PI3K-AKT pathways in the TN tumors by including the interacting non-coding microRNAs and circRNAs that, based on our results, are also essential members of these pathways. Figures 22 and 23 represent the MAPK and PI3K-AKT endogenous RNA networks with both protein-coding and non-coding RNAs that are likely to interact with each other and regulate expression of the central mRNAs, such as oncogenes PI3K, AKT, Ras, NFkB and tumor suppressor PTEN.



**Figure 22: MAPK endogenous RNA network.** This figure shows the network of interacting protein-coding and non-coding genes in the MAPK signaling pathway. The mRNAs and microRNAs reported by ReMix are represented in colors orange and blue, respectively. Intra-gene circRNAs are represented in green color. Oncogenes AKT, RAS, NF-κB, PI3K, ERK and MEK are shown to interact either directly or indirectly with the mRNA-microRNA-circRNA candidates.



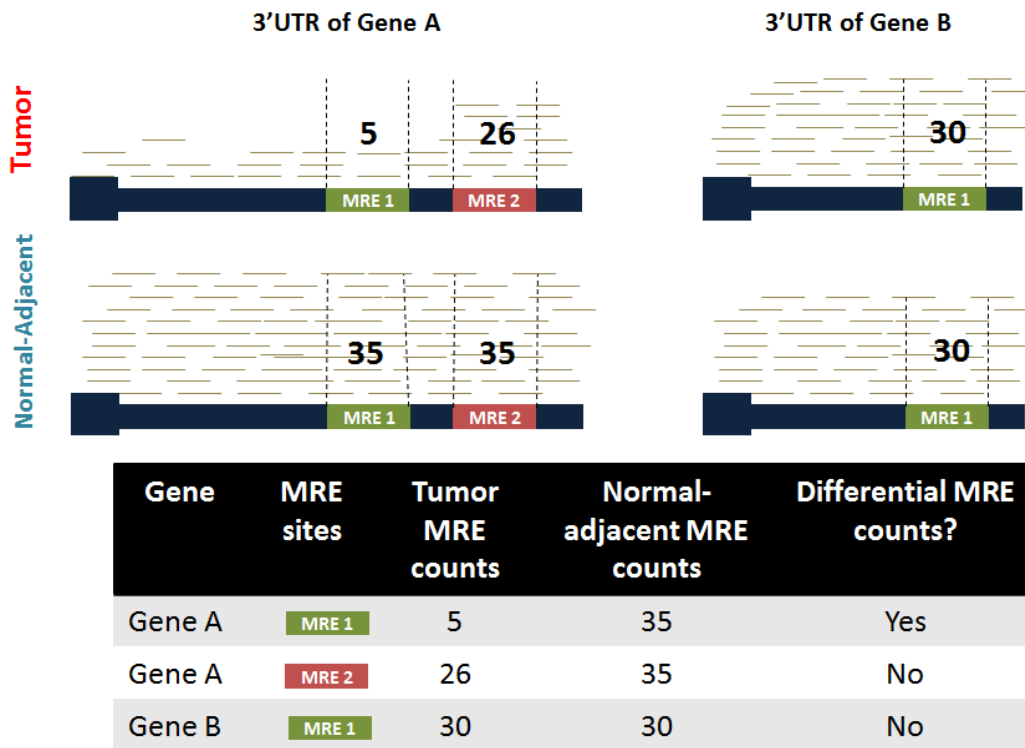
**Figure 23: PI3K-AKT endogenous RNA network.** This figure illustrates the network of interacting protein-coding and non-coding genes in the PI3K-AKT signaling pathway. The mRNAs and microRNAs reported by ReMix are represented in colors orange and blue, respectively. Intra-gene circRNAs are represented in green color. Oncogenes AKT, RAS, NF-kB and PI3K as well as tumor suppressor PTEN are shown to interact either directly or indirectly with the mRNA-microRNA-circRNA candidates.

At the end, by including the interacting mRNAs, microRNAs and circRNAs and further expanding the MAPK and PI3K-AKT signaling pathways, we believe that we have identified a set of protein-coding and non-coding RNAs that have the potential to influence and modulate the overall expression of these pathways. Based on further analysis using expression-correlation methods, we trust the possibility of finding candidate RNAs can be promising diagnostic targets towards improving current treatment modalities in TN breast cancer patients.

## 4.3 Methods

### 4.3.1 ReMix – a novel methodology to compute MRE frequency from RNA-Seq data

Here we present an innovative bioinformatics approach called ReMix, which was used to quantify MRE sites at the 3'UTR regions of mRNAs from RNA-Seq data. This methodology, called ReMix, uses reads aligned to the 3'UTR and scans them for evidence of any given MRE sequence. MRE sequences, which are complimentary to the seed sequences of microRNAs, are searched in the 3'UTR aligned reads of genes that are known to be associated with microRNAs (based on TargetScan – human version 7.0 [114]). A hypothetical example of this approach is illustrated in Figure 22. The reads aligned to 3'UTR regions of genes Gene A and Gene B are shown in Tumor and Normal-Adjacent samples. Gene A contains two MRE sites, and Gene B has one MRE site, with a common site (MRE1) in both genes. The number of reads mapped to these MRE sites are quantified for each gene and tabulated separately for Tumor and Normal-Adjacent samples. Later, MRE counts per gene are normalized and statically evaluated to identify differentially expressed MREs for downstream analysis.



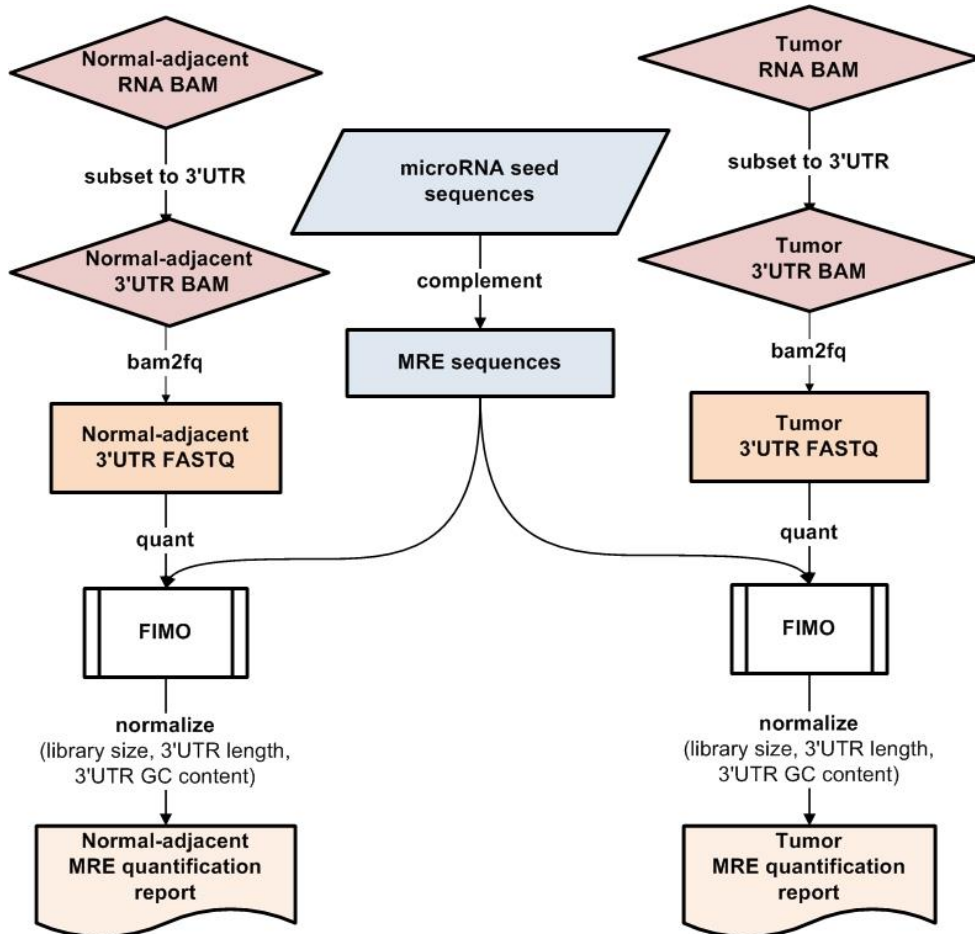
**Figure 24:** Hypothetical representation of MRE frequency counting using RNA-Seq data. The example shows a Tumor and Normal-Adjacent sample with reads mapped to the 3' UTRs of two genes, Gene A and Gene B that consist of 2 and 1 MRE sites respectively, with a common site (MRE 1). Reads that align with individual MRE sites (vertical dotted lines) are quantified. For every MRE site that belongs to a gene, the counts are then statistically evaluated between Tumor and Normal-Adjacent for evidence of differential frequency (as shown in the table).

#### 4.3.2 MRE frequency analysis from RNA-Seq data

Seed sequences for all the conserved microRNA families (n=320) were downloaded from TargetScanHuman 7.1, and the corresponding complementary MRE sequences were derived using in-house bioinformatics scripts. Figure 23 is



a flowchart representation of the bioinformatics methodology behind ReMlx, developed for MRE frequency analysis.



**Figure 25:** Flowchart representation of MRE frequency quantification from RNA-Seq BAM. The RNA-Seq BAM is first subset to 3'UTR regions of all genes, converted to FASTQ and then processed through FIMO to obtain raw MRE counts per microRNA for every target gene. The raw MRE counts are normalized to account for library size, 3'UTR length and 3'UTR GC content and individual Tumor and Normal-adjacent quantification reports are generated.

First, the RNA-Seq BAM files for both Tumor and Normal-Adjacent were subset to the 3'UTR regions for all genes using the SAMTools suite [58], thereby

obtaining one BAM per gene. Second, the newly obtained BAM files were converted into a FASTQ format so that for every gene, we had a collection of all reads that were mapped to their 3'UTR regions. BAM to FASTQ conversion was achieved using the bam2fastx module from Tophat [59]. Third, using the complementary seed (or MRE) sequences of individual microRNAs and the FASTQ files for corresponding genes that are known targets of these microRNAs, we quantified the occurrence of each MRE site by utilizing the motif searching bioinformatics tool called Find Individual Motif Occurrences (FIMO) [112]. Briefly, FIMO is an efficient and statistically rigorous software tool that scans nucleotide or protein sequences for the occurrence of a motif, or in this case, a MRE sequence. FIMO evaluates the occurrence of the MRE against a background frequency of nucleotides present in the hg19/NCBI 37.1 human reference genome and computes a log-likelihood ratio score. These scores are then converted to p-values, indicating whether the occurrence was a statically significant observation or not. The output from FIMO is a table indicating the number of occurrences per MRE and their associated p-values. MRE sites with p-value < 0.05 were selected from the FIMO output for downstream processing. Fourth, the raw MRE counts were normalized to account for the differences in library size/sequencing depths as well as the 3' UTR length and 3' UTR GC content per gene. Finally, for every gene and every conserved microRNA that targets the gene, the normalized MRE counts were reported in a tab-delimited

format for each gene-microRNA pair in the Tumor and Normal-Adjacent cases separately.

### **4.3.3 3'UTR definitions obtained from TargetScan**

Recently, Dr. David Bartel's group developed an improved quantitative model to predict canonical targeting of microRNAs to 3'UTR regions of mRNA [114]. A combination of 14 features in the model coupled with experimental approaches such as poly(A)-position profiling by sequencing called 3P-seq was used to define 3'UTR positions of genes in the transcriptome accurately. This data, available at the TargetScanHuman 7.1 database, is what was used for 3'UTR definitions of genes in the MRE analysis study.

### **4.3.4 RNA-Seq and microRNA-Seq data from TCGA**

The RNA-Seq binary alignment map (BAM) files and the microRNA Sequencing fastq files for the Triple Negative Tumor and Normal-Adjacent paired cases were downloaded from the TCGA Research Network (<http://cancergenome.nih.gov/>) using the National Cancer Institute (NCI) Genomic Data Commons (GDC) resource (<https://gdc.cancer.gov/>). The RNA-Seq BAM files were then converted to fastq files and aligned to the hg19/NCBI 37.1 human reference genome using the MAP-RSeq workflow [37]. The microRNA fastq files were aligned to the same human genome build using the CAP-miRSeq workflow [115]. The normalized

microRNA counts from the workflow were used to obtain the microRNA expression values in the Triple Negative samples. The unmapped reads from the RNA BAM files obtained from MAP-RSeq [37] were further processed through the Circ-Seq workflow [38] to identify and annotate stable and expressed circRNAs for all the Triple Negative Tumor and Normal-Adjacent samples.

#### **4.3.5 Statistical analyses on MRE sites and activated pathway identification**

Evaluation of MRE sites that represented distinct counts specific to TN tumors as opposed to ER+ and HER2+ subtypes as well as TN normal-adjacent pairs were obtained using the R package Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes. MRE sites significantly associated with TN tumors were selected based on a p-value cut-off  $< 0.05$ . Identification of the relevance and activation/inhibition status of pathways was evaluated using the R package called Signaling Pathway Impact Analysis (SPIA).

#### **4.3.6 Pathway analysis for canonical pathways**

Enriched canonical pathway analyses for 272 genes with differential MRE counts in TN breast tumors were identified using KEGG and Reactome functional

databases. Open source analysis toolkit WebGestalt [98] was used to for pathway identification by using the Gene set enrichment analysis (GSEA) option.

## 4.4 Discussion

In this study, we developed an innovative bioinformatics methodology called ReMlx that uses RNA-Seq data to identify and quantify microRNA binding sites or MREs at 3'UTR regions of genes. We applied this methodology to TCGA paired tumor and normal-adjacent breast cancer cases for TN, ER+, and HER2+ molecular subtypes and identified 614 MRE sites that had a distinct expression in TN tumors only. Majority of these sites had lower expression in TN tumors. Upon de-coupling, we found that the 614 MRE sites corresponded to 272 genes and 198 microRNAs. Canonical pathway analysis revealed that the top significant pathways in these TN tumors were MAPK and PI3K-AKT signaling cascades. About 4% of gene members in these pathways were represented by genes we obtained from the MRE analysis. Based on our previous work in circRNAs in TCGA breast cancer [38], we found a number of intra-gene circRNAs that were associated with genes in the MAPK and PI3K-AKT pathways. Based on our results from this study, we provide a list of potential mRNA-microRNA-circRNA candidates that are likely to interact with each other at the network level of MAPK and PI3K-AKT signaling cascades and could be possible targets for identification of novel biomarkers in TN tumors.

One of the limitations of this study is that while ReMix enables identification of candidate mRNA and microRNA players via MRE analysis using RNA-Seq data, this does not establish the fact that the identified microRNAs are indeed present and expressed in the particular disease, in this case, TN tumors. ReMix results only confirm that the sites on 3'UTR of mRNAs where microRNAs bind to, show distinct expression profiles in tumor and thus have the potential to be regulated by microRNAs in a disease-specific manner. To complement these results, microRNA expression profiles can be used to validate the existence of microRNAs and to check for expression correlation with the corresponding mRNA target(s) identified by ReMix.

The libraries for the TCGA RNA-Seq samples were prepared using Illumina's TruSeq library preparation protocol [110, 111]. One significant advantage of this protocol is that mRNAs are selected using poly-A enrichment. This enables pull-down of not just coding features of genes, but also 3' UTRs, to which poly-A tails are attached. While studies have utilized RNA-Seq to look into alternatively spliced 3'UTRs [116] and examine the impact of expressed variants present in the 3'UTR of genes [117, 118], no work has been done to date to extrapolate RNA-Seq data alone and obtain insights into the interactions between mRNAs and microRNAs at these 3' UTR regions.

Lately, there has been much focus in research to explore and identify therapeutic strategies to better treat Triple Negative breast cancer patients and improve their chances of survival. It has been shown that activation of the MAPK and PI3K-AKT pathways result in cancer cell proliferation and survival in the tumor [119]. Previous studies have shown these pathways to be highly prevalent in TN breast cancer as opposed to other breast cancer subtypes [120-122], thus supporting our findings. Studies have also shown that activation of MAPK [122-127] and PI3K-AKT [122, 128-134] pathways significantly correlate with tumor proliferation and disease progression in TN tumors. MAPK pathway is a sequentially activated cascade consisting of key genes such as Ras, Raf, MEK, and ERK. Activation of Ras leads to phosphorylation of Raf thereby advancing into activation of MEK and ERK downstream, finally resulting in tumor proliferation and cell survival. The PI3K-AKT pathway is yet another signaling cascade where kinase PI3K is the central driver of oncogenic transformation and plays a fundamental role in proliferation and tumor survival. PI3K activates oncogene AKT that modulates downstream signaling pathways such as mTOR. The end result is inhibition of apoptosis and increased cell proliferation.

Pre-clinical experiments have shown that MAPK and PI3K-AKT pathways have significant cross-talk and that inhibition of one cascade activates the other, and vice versa [119]. Thus, pre-clinical trials are underway that focus on co-targeting both pathways to improve clinical outcomes [119]. At present, [clinicaltrials.gov](http://clinicaltrials.gov)

lists identifiers NCT01623349 (active), NCT03337724 (not yet recruiting), NCT03218826 (not yet recruiting) and NCT01629615 (completed) as trials that are investigating the effect of various PI3K drug inhibitors in Triple Negative patients. We believe that upon further evaluation and validation of our current mRNA-microRNA-circRNA results in this study, there is a potential to identify novel biomarkers, especially circRNA and microRNA candidates that interact with mRNA targets and regulate their gene expression in MAPK and PI3K-AKT pathways.



## Chapter 5: Conclusions and Discussion

RNA-Seq is a significant breakthrough in Next Generation Sequencing technology and has become the standard in bioinformatics for analysis of the transcriptome. The vast wealth of information offered by RNA-Seq helps uncover not only numerous features of protein-coding regions but also non-coding regions that are expressed in the transcriptome. There are many stand-alone bioinformatics packages which analyze different aspects of the transcriptome. However, they cannot be plugged together quickly, and they do not provide a complete, integrated picture. In addition, the field of bioinformatics lacks simple-to-use workflows that can comprehensively analyze several features of the transcriptome together, and report in an integrated fashion. The first half of this dissertation offers bioinformatics solutions to address the above challenges.

**Chapter 1** is an introduction to the various RNA types in the human transcriptome that are studied in this dissertation, namely – mRNAs (protein-coding), microRNAs and circRNAs (both non-coding). The unique interaction between these RNA types as well as the possible implications of these interactions in the etiology of diseases such as breast cancer is outlined in this chapter. Existing bioinformatics challenges are identified, and the motivation to address these is established by listing the three main specific aims that are pursued in this dissertation.

**Chapter 2** presents a comprehensive bioinformatics workflow called MAP-RSeq [37] that was developed to both analyze and obtain several features of protein-coding regions in the transcriptome. MAP-RSeq consists of six major modules such as alignment of reads, quality assessment of reads, gene expression assessment and exon read counting, identification of expressed single nucleotide variants (SNVs), detection of fusion transcripts, and summarization of transcriptome data in a final report.

**Chapter 3** presents a bioinformatics workflow called Circ-Seq [38] that was developed to characterize circRNAs, which are newly discovered and highly stable forms of non-coding RNAs with diverse biological functions. Circ-Seq consists of five major modules such as alignment of unmapped reads, identification of circRNA junctions, application of circRNA specific filters and extensive genomic annotation of expressed circRNA candidates, all summarized in a final report.

Circ-Seq can be used seamlessly after MAP-RSeq, and at the end of both workflows, researchers can quickly and efficiently acquire information on various perspectives of protein-coding mRNAs and non-coding circRNAs without spending time and effort in dealing with multiple stand-alone packages and integrating results across them.

In this dissertation, MAP-RSeq and Circ-Seq workflows were applied to the largest cohort of breast cancer (n=885 samples) from The Cancer Genome Atlas (TCGA) as well as breast cancer and breast normal cell lines. MAP-RSeq results were used to obtain gene expression profiles across all three molecular subtypes – ER+, HER2+, and TN, of breast cancer. By applying the Circ-Seq workflow and analyzing the results, we reported for the first time in breast cancer research, the identification and annotation of a 7kb long circRNA, which was experimentally validated in the MCF7 breast cancer cell line using qRT-PCR and Sanger sequencing. In addition, comprehensive analysis of circRNAs in TCGA ER+ breast cancer subtype suggested that circRNA frequency may be a marker for cell proliferation in breast cancer.

Complex interactions between mRNAs, circRNAs, and microRNAs can greatly influence post-transcriptional activity in the cell. For heterogeneous diseases such as breast cancer, it is of paramount value to understand such interactions and elucidate the key players that cause ultimate changes in target gene expression in the disease. It is a well-known fact that microRNAs interact with mRNAs through MRE binding sites, located on the 3' UTR of genes, and regulate gene expression. Studies have looked into mRNA-microRNA interactions using their expression profiles, alternative polyadenylation of 3'UTRs, SNPs along MRE sites, etc. However, no study has investigated the use of RNA-Seq data

alone to obtain MRE frequencies and gather insights about MREs that show evidence of being differentially expressed in tumors until now, and this is being addressed in this dissertation.

**Chapter 4** of this dissertation presents a novel bioinformatics approach called ReMlx that uses TCGA RNA-Seq data of tumor and normal-adjacent pairs and quantifies known MRE sites by screening 3'UTR regions across the whole transcriptome. We focused on the TN breast cancer subtype and by using definitive statistical measures, identified a set of MREs that showed distinct expression in TN tumors, when compared to TN normal-adjacent, ER+, and HER2+ subtypes. MRE results from ReMlx also highlighted a) mRNA genes in TN tumors that contain the differential MREs, and b) candidate microRNAs that have the potential to regulate these genes. Further analysis showed that the identified genes belonged to MAPK and PI3K-AKT pathway cascades. These genes have direct or indirect interactions with some of the key cancer genes in these pathways, such as AKT and Ras oncogenes, which are activated, and tumor suppressors such as PTEN and NFkB, which are repressed in TN subtype. Finally, we overlaid the TN tumor circRNAs that were obtained from the Circ-Seq workflow to these pathways and found that a subset of genes, both from the ReMlx analysis as well as gene members of these pathways, contain expressed circRNA transcripts. Here, we have identified a set of mRNA-microRNA-circRNA candidates that have the potential to interact with each other

in a TN tumor-specific manner and regulate some of the vital cancer genes within the MAPK and PI3K-AKT pathways.

There are several directions in which work from this dissertation can be pursued forward. For example, the mRNA and microRNA candidates identified in chapter 4 using the ReMix approach need to be further investigated to elucidate how their changes in expression can transmit changes to their neighboring gene partners and how this eventually affects the expression of cancer-relevant genes. At present, there are no targeted treatment options available for TN breast cancer. Identification of novel biomarkers for this disease type is the need of the hour, and we hope that the results from this dissertation can help shed light towards identifying new targets for TN breast cancer. Studies have shown that PIK3CA mutations are found in approximately 20% of breast cancers [122, 135, 136]. It is hypothesized that PIK3CA may have a role in tumor proliferation in TN breast cancer. However, the impact of mutated PIK3CA on the TN subtype has not been established [135, 136]. Likewise, mutations in AKT1 are reported in about 8% of breast cancers, which lead to higher kinase activity than the wild-type AKT1 [137], but apparently, this has to be investigated in TN breast cancer. MAP-RSeq reports genomic variants such as SNVs and gene fusions as part of the workflow output. These reports can be further explored to look into MAPK and PI3K-AKT pathways to identify any aberrations in the TN breast cancer cases. Further, these findings can be correlated with the ReMix results to check

whether mutations lead to increased or decreased MRE sites in the affected genes in TN tumors.

Another prospective approach for ReMix algorithm in chapter 4 is to analyze the ER+ and HER2+ breast cancer subtypes. We have already shown from our Circ-Seq results for the TCGA ER+ subtype that Luminal A ER+ tumors (which are less proliferative) appear to have a higher number of circRNAs compared to the more proliferative Luminal B ER+ tumors, indicating that presence of more circRNAs tends to control tumor growth. It is not known how this regulatory mechanism works, but one could hypothesize that circRNAs in Luminal A tumors aid with the sustained activity of tumor suppressor genes by acting as microRNA sponges, thereby sequestering microRNAs towards them, which otherwise would bind to tumor suppressor genes and induce repression. Our ReMix methodology can be used to identify candidate mRNA-microRNA pairs that show differential MRE sites in Luminal A and B ER+ tumors and thereafter, validate whether the hypothesis above holds true. Similar analysis can also be performed with ER+/HER2+ and ER-/HER2+ in breast cancer and other cancers as well.

The work in this dissertation was focused on using RNA-Seq data alone to characterize various features of protein-coding/mRNAs and non-coding circRNAs, and also to investigate regulatory elements along 3'UTR regions of mRNAs to assess their potential impact on mRNA-microRNA-circRNA

interactions in breast cancer. One could extend this work by including data from other sequencing applications such as whole-exome (WES) or whole-genome (WGS) from DNA-Seq and corroborate the results presented in this dissertation. For example, a commonly-used approach is to use WES to verify SNVs identified in RNA-Seq. Using SNVs results that are reported by MAPR-Seq for the TCGA breast samples, one could integrate SNVs from WES and subset to candidates reported in both sequencing types. These SNVs could then be further investigated to identify mutation signatures that are specific to the three breast cancer subtypes. Another approach is to use WES that includes 3'UTR regions. Studies have shown that alternative polyadenylation (APA) can lead to either short or long 3'UTRs in TN tumors [138]. Thus, one could inspect whether ReMix gene candidates, with differential MRE sites in TN breast cancer, show differences due to dissimilar 3'UTR sizes in tumor and normal-adjacent cases. To investigate noncoding RNAs, WGS offers a great validation platform. For example, unmapped reads from WGS can be used to confirm the non-coding circRNA candidates found by Circ-Seq, by checking for the presence of 3'-5' reads that were identified using RNA-Seq data.

This dissertation has taken a deep dive into non-coding circRNAs and understanding their associations with breast cancer subtypes. However, there are other non-coding RNAs such as lincRNAs that can also be brought into the picture. As known, lincRNAs are widely present in the human genome, exhibit

tissue-specific expression and importantly, have strong regulatory implications on target mRNAs [139]. Additionally, lincRNAs can also contain MRE sites along their 3'UTR regions and therefore compete with mRNAs and circRNAs for microRNA binding [140-142]. Thus, work from this dissertation can be extended to include lincRNAs, and potentially other non-coding RNAs like pseudogenes as well, to investigate the complex and competitive biological interactions in different types of RNAs from tumor and normal cell environments.

In conclusion, the future of bioinformatics research is promising. There is a continued demand for bioinformatics advancements, especially towards resolving deadly diseases such as cancer and to enable the research community in deconvoluting the complex biological networks in the transcriptome. We need to understand the intricate interactions between various RNA types that lead to the ultimate transformation from normal to tumor cells, and bioinformatics research can uncover such answers.



## Bibliography

1. Neafsey DE, Palumbi SR: **Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes.** *Genome research* 2003, **13**:821-830.
2. Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM: **The mode and tempo of genome size evolution in eukaryotes.** *Genome research* 2007, **17**:594-601.
3. Redi CA, Capanna E: **Genome size evolution: sizing mammalian genomes.** *Cytogenetic and genome research* 2012, **137**:97-112.
4. Moran LA: **Genome Size, Complexity, and the C-Value Paradox.** 2007.
5. Keinath MC, Timoshevskiy VA, Timoshevskaya NY, Tsonis PA, Voss SR, Smith JJ: **Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing.** *Scientific reports* 2015, **5**:16413.
6. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:19428-19433.
7. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: **Annotating non-coding regions of the genome.** *Nature reviews Genetics* 2010, **11**:559-571.
8. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: **Cross-species sequence comparisons: a review of methods and available resources.** *Genome research* 2003, **13**:1-12.
9. **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57-74.
10. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.** *Nature* 2012, **489**:101-108.
11. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome research* 2012, **22**:1760-1774.
12. Ponting CP, Oliver PL, Reik W: **Evolution and functions of long noncoding RNAs.** *Cell* 2009, **136**:629-641.
13. Malone CD, Hannon GJ: **Small RNAs as guardians of the genome.** *Cell* 2009, **136**:656-668.
14. Carthew RW, Sontheimer EJ: **Origins and Mechanisms of miRNAs and siRNAs.** *Cell* 2009, **136**:642-655.
15. Balakirev ES, Ayala FJ: **Pseudogenes: are they "junk" or functional DNA?** *Annual review of genetics* 2003, **37**:123-151.
16. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al: **Circular RNAs are a**

- large class of animal RNAs with regulatory potency.** *Nature* 2013, **495**:333-338.
17. Pratt AJ, MacRae IJ: **The RNA-induced silencing complex: a versatile gene-silencing machine.** *The Journal of biological chemistry* 2009, **284**:17897-17901.
  18. Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, Hanan M, Wyler E, Perez-Hernandez D, Ramberger E, et al: **Translation of CircRNAs.** *Molecular cell* 2017, **66**:9-21 e27.
  19. Mullany LE, Herrick JS, Wolff RK, Slattery ML: **MicroRNA Seed Region Length Impact on Target Messenger RNA Expression and Survival in Colorectal Cancer.** *PloS one* 2016, **11**:e0154177.
  20. Price C, Chen J: **MicroRNAs in Cancer Biology and Therapy: Current Status and Perspectives.** *Genes & diseases* 2014, **1**:53-63.
  21. Buermans HP, den Dunnen JT: **Next generation sequencing technology: Advances and applications.** *Biochimica et biophysica acta* 2014, **1842**:1932-1941.
  22. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities.** *Nature reviews Genetics* 2011, **12**:87-98.
  23. Creighton CJ, Hernandez-Herrera A, Jacobsen A, Levine DA, Mankoo P, Schultz N, Du Y, Zhang Y, Larsson E, Sheridan R, et al: **Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma.** *PloS one* 2012, **7**:e34546.
  24. Fackler MJ, Umbricht CB, Williams D, Argani P, Cruz LA, Merino VF, Teo WW, Zhang Z, Huang P, Visvanathan K, et al: **Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence.** *Cancer research* 2011, **71**:6195-6207.
  25. Farazi TA, Horlings HM, Ten Hoeve JJ, Mihailovic A, Halfwerk H, Morozov P, Brown M, Hafner M, Reyat F, van Kouwenhove M, et al: **MicroRNA sequence and expression analysis in breast tumors by deep sequencing.** *Cancer research* 2011, **71**:4443-4453.
  26. Guo L, Zhao Y, Yang S, Zhang H, Chen F: **Integrative analysis of miRNA-mRNA and miRNA-miRNA interactions.** *BioMed research international* 2014, **2014**:907420.
  27. Stricker TP, Brown CD, Bandlamudi C, McNerney M, Kittler R, Montoya V, Peterson A, Grossman R, White KP: **Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression.** *PLoS genetics* 2017, **13**:e1006589.
  28. Volinia S, Croce CM: **Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer.** *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:7413-7417.
  29. Vu TN, Pramana S, Calza S, Suo C, Lee D, Pawitan Y: **Comprehensive landscape of subtype-specific coding and non-coding RNA transcripts in breast cancer.** *Oncotarget* 2016, **7**:68851-68863.

30. Zhang Y, Li Y, Wang Q, Zhang X, Wang D, Tang HC, Meng X, Ding X: **Identification of an lncRNAmiRNAmRNA interaction mechanism in breast cancer based on bioinformatic analysis.** *Molecular medicine reports* 2017, **16**:5113-5120.
31. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.** *Genome biology* 2011, **12**:R72.
32. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30**:923-930.
33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome research* 2010, **20**:1297-1303.
34. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841-842.
35. Sun K, Zhao Y, Wang H, Sun H: **Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs)-- application in human skeletal muscle cells.** *PloS one* 2014, **9**:e84500.
36. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Briefings in bioinformatics* 2013, **14**:178-192.
37. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, Nie J, Tang X, Baheti S, Doughty JB, et al: **MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing.** *BMC bioinformatics* 2014, **15**:224.
38. Nair AA, Niu N, Tang X, Thompson KJ, Wang L, Kocher JP, Subramanian S, Kalari KR: **Circular RNAs and their associations with breast cancer subtypes.** *Oncotarget* 2016, **7**:80967-80979.
39. Barrett CL, Schwab RB, Jung H, Crain B, Goff DJ, Jamieson CH, Thistlethwaite PA, Harismendy O, Carson DA, Frazer KA: **Transcriptome sequencing of tumor subpopulations reveals a spectrum of therapeutic options for squamous cell lung cancer.** *PloS one* 2013, **8**:e58714.
40. Chen Y, Souaiaia T, Chen T: **PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds.** *Bioinformatics* 2009, **25**:2514-2521.
41. Head SR, Mondala T, Gelbart T, Ordoukhanian P, Chappel R, Hernandez G, Salomon DR: **RNA purification and expression analysis using microarrays and RNA deep sequencing.** *Methods in molecular biology* 2013, **1034**:385-403.
42. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139-140.
43. 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**:e178.

44. Goncalves A, Tikhonov A, Brazma A, Kapushesky M: **A pipeline for RNA-seq data processing and quality assessment.** *Bioinformatics* 2011, **27**:867-869.
45. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.** *Bioinformatics* 2011, **27**:281-283.
46. Qi J, Zhao F, Buboltz A, Schuster SC: **inGAP: an integrated next-generation genome analysis pipeline.** *Bioinformatics* 2010, **26**:127-129.
47. Wang Y, Mehta G, Mayani R, Lu J, Souaiaia T, Chen Y, Clark A, Yoon HJ, Wan L, Evgrafov OV, et al: **RseqFlow: workflows for RNA-Seq data analysis.** *Bioinformatics* 2011, **27**:2598-2600.
48. Kalari KR, Nair, A. A: **MAP-RSeq website.** *MC Bioinformatics* 2014.
49. webpage VBd: **Virtual Box download webpage.**
50. webpage C: **CGHub webpage.**
51. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome research* 2002, **12**:656-664.
52. 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**:R25.
53. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome research* 2009, **19**:1639-1645.
54. website F: **FASTQC website.**
55. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31**:166-169.
56. webpage PT: **Picard Tools webpage.**
57. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments.** *Bioinformatics* 2012, **28**:2184-2185.
58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.
59. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105-1111.
60. Egan JB, Barrett MT, Champion MD, Middha S, Lenkiewicz E, Evers L, Francis P, Schmidt J, Shi CX, Van Wier S, et al: **Whole genome analyses of a well-differentiated liposarcoma reveals novel SYT1 and DDR2 rearrangements.** *PloS one* 2014, **9**:e87113.
61. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, Jen J, Eckloff BW, Kalari KR, Thompson KJ, et al: **Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors.** *PloS one* 2013, **8**:e81925.
62. Sakuma T, Davila JI, Malcolm JA, Kocher JP, Tonne JM, Ikeda Y: **Murine leukemia virus uses NXF1 for nuclear export of spliced and unspliced viral transcripts.** *Journal of virology* 2014, **88**:4069-4082.
63. annotation Cia: **Cufflink index and annotation.**

64. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)**. *Bioinformatics* 2011, **27**:2518-2528.
65. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC bioinformatics* 2010, **11**:422.
66. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome biology* 2010, **11**:R106.
67. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.
68. Sonesson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC bioinformatics* 2013, **14**:91.
69. Seyednasrollah F, Laiho A, Elo LL: **Comparison of software packages for detecting differential expression in RNA-seq studies**. *Briefings in bioinformatics* 2015, **16**:59-70.
70. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D: **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data**. *Genome biology* 2013, **14**:R95.
71. Guo JU, Agarwal V, Guo H, Bartel DP: **Expanded identification and characterization of mammalian circular RNAs**. *Genome biology* 2014, **15**:409.
72. 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**:e30733.
73. 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* 2013, **19**:141-157.
74. Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A: **Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing**. *Nucleic acids research* 2006, **34**:e63.
75. 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**:3852-3856.
76. Diener TO: **Viroids: structure and function**. *Science* 1979, **205**:859-866.
77. 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**:1019-1030.
78. Pasman Z, Been MD, Garcia-Blanco MA: **Exon circularization in mammalian nuclear extracts**. *RNA* 1996, **2**:603-610.
79. Chao CW, Chan DC, Kuo A, Leder P: **The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis**. *Molecular medicine* 1998, **4**:614-628.

80. 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**:55-66.
81. Wang PL, Bao Y, Yee MC, Barrett SP, Hogan GJ, Olsen MN, Dinneny JR, Brown PO, Salzman J: **Circular RNA is expressed across the eukaryotic tree of life.** *PloS one* 2014, **9**:e90859.
82. Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO: **Cell-type specific features of circular RNA expression.** *PLoS genetics* 2013, **9**:e1003777.
83. Jeck WR, Sharpless NE: **Detecting and characterizing circular RNAs.** *Nature biotechnology* 2014, **32**:453-461.
84. Lukiw WJ: **Circular RNA (circRNA) in Alzheimer's disease (AD).** *Frontiers in genetics* 2013, **4**:307.
85. Valdmanis PN, Kay MA: **The expanding repertoire of circular RNAs.** *Molecular therapy : the journal of the American Society of Gene Therapy* 2013, **21**:1112-1114.
86. Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al: **Exon-intron circular RNAs regulate transcription in the nucleus.** *Nature structural & molecular biology* 2015, **22**:256-264.
87. 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**:384-388.
88. Bahn JH, Zhang Q, Li F, Chan TM, Lin X, Kim Y, Wong DT, Xiao X: **The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva.** *Clinical chemistry* 2015, **61**:221-230.
89. Li Y, Zheng Q, Bao C, Li S, Guo W, Zhao J, Chen D, Gu J, He X, Huang S: **Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis.** *Cell research* 2015, **25**:981-984.
90. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH: **starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data.** *Nucleic acids research* 2011, **39**:D202-209.
91. 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**:8057.
92. 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**:283.
93. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, Davies SR, Snider J, Stijleman IJ, Reed J, et al: **A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer.** *Clinical cancer research : an official journal of the American Association for Cancer Research* 2010, **16**:5222-5232.

94. Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, Carr JM, Khrebtukova I, Luo S, Zhang L, et al: **Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing.** *PloS one* 2011, **6**:e17490.
95. Dai X, Li Y, Bai Z, Tang XQ: **Molecular portraits revealing the heterogeneity of breast tumor subtypes defined using immunohistochemistry markers.** *Scientific reports* 2015, **5**:14499.
96. Pouladi N, Cowper-Sallari R, Moore JH: **Combining functional genomics strategies identifies modular heterogeneity of breast cancer intrinsic subtypes.** *BioData mining* 2014, **7**:27.
97. Martin M, Prat A, Rodriguez-Lescure A, Caballero R, Ebbert MT, Munarriz B, Ruiz-Borrego M, Bastien RR, Crespo C, Davis C, et al: **PAM50 proliferation score as a predictor of weekly paclitaxel benefit in breast cancer.** *Breast cancer research and treatment* 2013, **138**:457-466.
98. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic acids research* 2005, **33**:W741-748.
99. Aherne ST, Madden SF, Hughes DJ, Pardini B, Naccarati A, Levy M, Vodicka P, Neary P, Dowling P, Clynes M: **Circulating miRNAs miR-34a and miR-150 associated with colorectal cancer progression.** *BMC cancer* 2015, **15**:329.
100. Hoffman Y, Bublik DR, Pilpel Y, Oren M: **miR-661 downregulates both Mdm2 and Mdm4 to activate p53.** *Cell death and differentiation* 2014, **21**:302-309.
101. Illumina: **Illumina TrueSeq RNA preparation.**
102. Prueitt RL, Yi M, Hudson RS, Wallace TA, Howe TM, Yfantis HG, Lee DH, Stephens RM, Liu CG, Calin GA, et al: **Expression of microRNAs and protein-coding genes associated with perineural invasion in prostate cancer.** *The Prostate* 2008, **68**:1152-1164.
103. Schetter AJ, Leung SY, Sohn JJ, Zanetti KA, Bowman ED, Yanaihara N, Yuen ST, Chan TL, Kwong DL, Au GK, et al: **MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma.** *JAMA* 2008, **299**:425-436.
104. Lee S, Jiang X: **Modeling miRNA-mRNA interactions that cause phenotypic abnormality in breast cancer patients.** *PloS one* 2017, **12**:e0182666.
105. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.** *Nature* 2005, **433**:769-773.
106. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nature genetics* 2005, **37**:495-500.
107. Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C: **Analysis of microRNA-target interactions across diverse cancer types.** *Nature structural & molecular biology* 2013, **20**:1325-1332.

108. Han S, Kim D, Shivakumar M, Lee YJ, Garg T, Miller JE, Kim JH, Lee Y: **The effects of alternative splicing on miRNA binding sites in bladder cancer.** *PLoS one* 2018, **13**:e0190708.
109. Pelletier C, Weidhaas JB: **MicroRNA binding site polymorphisms as biomarkers of cancer risk.** *Expert review of molecular diagnostics* 2010, **10**:817-829.
110. **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**:61-70.
111. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, et al: **Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer.** *Cell* 2015, **163**:506-519.
112. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics* 2011, **27**:1017-1018.
113. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:15545-15550.
114. Agarwal V, Bell GW, Nam JW, Bartel DP: **Predicting effective microRNA target sites in mammalian mRNAs.** *eLife* 2015, **4**.
115. Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, Kocher JP: **CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data.** *BMC genomics* 2014, **15**:423.
116. Le Pera L, Mazzapioda M, Tramontano A: **3USS: a web server for detecting alternative 3'UTRs from RNA-seq experiments.** *Bioinformatics* 2015, **31**:1845-1847.
117. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets.** *PLoS biology* 2004, **2**:e363.
118. Skeeles LE, Fleming JL, Mahler KL, Toland AE: **The impact of 3'UTR variants on differential expression of candidate cancer susceptibility genes.** *PLoS one* 2013, **8**:e58609.
119. Saini K, J Piccart-Gebhart M: *Dual targeting of the PI3K and MAPK pathways in breast cancer.* 2010.
120. Balko JM, Miller TW, Morrison MM, Hutchinson K, Young C, Rinehart C, Sanchez V, Jee D, Polyak K, Prat A, et al: **The receptor tyrosine kinase ErbB3 maintains the balance between luminal and basal breast epithelium.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:221-226.
121. Hoeflich KP, O'Brien C, Boyd Z, Cavet G, Guerrero S, Jung K, Januario T, Savage H, Punnoose E, Truong T, et al: **In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models.** *Clinical cancer research : an official journal of the American Association for Cancer Research* 2009, **15**:4649-4664.



122. Hashimoto K, Tsuda H, Koizumi F, Shimizu C, Yonemori K, Ando M, Kodaira M, Yunokawa M, Fujiwara Y, Tamura K: **Activated PI3K/AKT and MAPK pathways are potential good prognostic markers in node-positive, triple-negative breast cancer.** *Annals of oncology : official journal of the European Society for Medical Oncology* 2014, **25**:1973-1979.
123. Loi S, Dushyanthen S, Beavis PA, Salgado R, Denkert C, Savas P, Combs S, Rimm DL, Giltane JM, Estrada MV, et al: **RAS/MAPK Activation Is Associated with Reduced Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: Therapeutic Cooperation Between MEK and PD-1/PD-L1 Immune Checkpoint Inhibitors.** *Clinical cancer research : an official journal of the American Association for Cancer Research* 2016, **22**:1499-1509.
124. Qi X, Yin N, Ma S, Lepp A, Tang J, Jing W, Johnson B, Dwinell MB, Chitambar CR, Chen G: **p38gamma MAPK Is a Therapeutic Target for Triple-Negative Breast Cancer by Stimulation of Cancer Stem-Like Cell Expansion.** *Stem cells* 2015, **33**:2738-2747.
125. Gholami S, Chen CH, Gao S, Lou E, Fujisawa S, Carson J, Nnoli JE, Chou TC, Bromberg J, Fong Y: **Role of MAPK in oncolytic herpes viral therapy in triple-negative breast cancer.** *Cancer gene therapy* 2014, **21**:283-289.
126. Giltane JM, Balko JM: **Rationale for targeting the Ras/MAPK pathway in triple-negative breast cancer.** *Discovery medicine* 2014, **17**:275-283.
127. Eralp Y, Derin D, Ozluk Y, Yavuz E, Guney N, Saip P, Muslumanoglu M, Igci A, Kucucuk S, Dincer M, et al: **MAPK overexpression is associated with anthracycline resistance and increased risk for recurrence in patients with triple-negative breast cancer.** *Annals of oncology : official journal of the European Society for Medical Oncology* 2008, **19**:669-674.
128. Massihnia D, Galvano A, Fanale D, Perez A, Castiglia M, Incorvaia L, Listi A, Rizzo S, Cicero G, Bazan V, et al: **Triple negative breast cancer: shedding light onto the role of pi3k/akt/mtor pathway.** *Oncotarget* 2016, **7**:60712-60722.
129. de Lint K, Poell JB, Soueidan H, Jastrzebski K, Vidal Rodriguez J, Lieftink C, Wessels LF, Beijersbergen RL: **Sensitizing Triple-Negative Breast Cancer to PI3K Inhibition by Cotargeting IGF1R.** *Molecular cancer therapeutics* 2016, **15**:1545-1556.
130. Kriegsmann M, Endris V, Wolf T, Pfarr N, Stenzinger A, Loibl S, Denkert C, Schneeweiss A, Budczies J, Sinn P, Weichert W: **Mutational profiles in triple-negative breast cancer defined by ultradeep multigene sequencing show high rates of PI3K pathway alterations and clinically relevant entity subgroup specific differences.** *Oncotarget* 2014, **5**:9952-9965.
131. Lehmann BD, Bauer JA, Schafer JM, Pendleton CS, Tang L, Johnson KC, Chen X, Balko JM, Gomez H, Arteaga CL, et al: **PIK3CA mutations in androgen receptor-positive triple negative breast cancer confer sensitivity to the combination of PI3K and androgen receptor inhibitors.** *Breast cancer research : BCR* 2014, **16**:406.
132. De P, Sun Y, Carlson JH, Friedman LS, Leyland-Jones BR, Dey N: **Doubling down on the PI3K-AKT-mTOR pathway enhances the antitumor efficacy of**

- PARP inhibitor in triple negative breast cancer model beyond BRCA-ness.** *Neoplasia* 2014, **16**:43-72.
133. Yi YW, Hong W, Kang HJ, Kim HJ, Zhao W, Wang A, Seong YS, Bae I: **Inhibition of the PI3K/AKT pathway potentiates cytotoxicity of EGFR kinase inhibitors in triple-negative breast cancer cells.** *Journal of cellular and molecular medicine* 2013, **17**:648-656.
134. Ibrahim YH, Garcia-Garcia C, Serra V, He L, Torres-Lockhart K, Prat A, Anton P, Cozar P, Guzman M, Grueso J, et al: **PI3K inhibition impairs BRCA1/2 expression and sensitizes BRCA-proficient triple-negative breast cancer to PARP inhibition.** *Cancer discovery* 2012, **2**:1036-1047.
135. Kalinsky K, Jacks LM, Heguy A, Patil S, Drobnjak M, Bhanot UK, Hedvat CV, Traina TA, Solit D, Gerald W, Moynahan ME: **PIK3CA mutation associates with improved outcome in breast cancer.** *Clinical cancer research : an official journal of the American Association for Cancer Research* 2009, **15**:5049-5059.
136. Berns K, Horlings HM, Hennessy BT, Madiredjo M, Hijmans EM, Beelen K, Linn SC, Gonzalez-Angulo AM, Stemke-Hale K, Hauptmann M, et al: **A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer.** *Cancer cell* 2007, **12**:395-402.
137. Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, Hostetter G, Boguslawski S, Moses TY, Savage S, et al: **A transforming mutation in the pleckstrin homology domain of AKT1 in cancer.** *Nature* 2007, **448**:439-444.
138. Miles WO, Lembo A, Volorio A, Brachtel E, Tian B, Sgroi D, Provero P, Dyson N: **Alternative Polyadenylation in Triple-Negative Breast Tumors Allows NRAS and c-JUN to Bypass PUMILIO Posttranscriptional Regulation.** *Cancer research* 2016, **76**:7231-7241.
139. Wang Q, Gao S, Li H, Lv M, Lu C: **Long noncoding RNAs (lncRNAs) in triple negative breast cancer.** *Journal of cellular physiology* 2017, **232**:3226-3233.
140. Eades G, Wolfson B, Zhang Y, Li Q, Yao Y, Zhou Q: **lincRNA-RoR and miR-145 regulate invasion in triple-negative breast cancer via targeting ARF6.** *Molecular cancer research : MCR* 2015, **13**:330-338.
141. Xia T, Liao Q, Jiang X, Shao Y, Xiao B, Xi Y, Guo J: **Long noncoding RNA associated-competing endogenous RNAs in gastric cancer.** *Scientific reports* 2014, **4**:6088.
142. Zhao Y, Wang H, Wu C, Yan M, Wu H, Wang J, Yang X, Shao Q: **Construction and investigation of lncRNA-associated ceRNA regulatory network in papillary thyroid cancer.** *Oncology reports* 2018, **39**:1197-1206.

# **Appendix**

## **6.1 Permissions**

This section contains Creative Commons license agreement and journal permissions obtained from BMC Bioinformatics and Oncotarget to reproduce published articles that are part of this dissertation. Snapshots of the published articles are also provided.

Dear Dr. Nair,

Thank you for contacting BioMed Central.

The open access articles published in BioMed Central's journals are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be re-used in any way, subject only to proper attribution (which, in an academic context, usually means citation).

The re-use rights enshrined in our license agreement (<http://www.biomedcentral.com/about/policies/license-agreement>) include the right for anyone to produce printed copies themselves, without formal permission or payment of permission fees. As a courtesy, however, anyone wishing to reproduce large quantities of an open access article (250+) should inform the copyright holder and we suggest a contribution in support of open access publication (see suggested contributions at <http://www.biomedcentral.com/about/policies/reprints-and-permissions/suggested-contributions>).

Please note that the following journals have published a small number of articles that, while freely accessible, are not open access as outlined above: Alzheimer's Research & Therapy, Arthritis Research & Therapy, Breast Cancer Research, Critical Care, Genome Biology, Genome Medicine, Stem Cell Research & Therapy.

You will be able to find details about these articles at <http://www.biomedcentral.com/about/policies/reprints-and-permissions>

If you have any questions, please do not hesitate to contact me.

Best wishes,

---

**Ricardo Sison Jr.**

Global Open Research Support Executive  
Open Research Group

**Springer Nature**

236 Gray's Inn Road, London WC1X 8HB, UK

T +44 (0)203 192 2009

F +44 (0)203 192 2010

[Ricardo.SisonJr@springernature.com](mailto:Ricardo.SisonJr@springernature.com)

[www.springernature.com](http://www.springernature.com)

---

Springer Nature is one of the world's leading global research, educational and professional publishers, created in 2015 through the combination of Nature Publishing Group, Palgrave Macmillan, Macmillan Education and Springer Science+Business Media.

---

Springer-Verlag London Ltd.

Registered Office: 236 Gray's Inn Road / London WC1X 8HB /

Registered in England

No. 1738860 / VAT Registration No. GB 823826326

**Figure 26:** Permission from BMC Bioinformatics to reproduce MAP-RSeq publication

SOFTWARE

Open Access

## MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing

Krishna R Kalari<sup>1†</sup>, Asha A Nair<sup>1†</sup>, Jaysheel D Bhavsar<sup>1</sup>, Daniel R O'Brien<sup>1</sup>, Jaime I Davila<sup>1</sup>, Matthew A Bockol<sup>1</sup>, Jinfu Nie<sup>1</sup>, Xiaojia Tang<sup>1</sup>, Saurabh Baheti<sup>1</sup>, Jay B Doughty<sup>1</sup>, Sumit Middha<sup>1</sup>, Hugues Sicotte<sup>1</sup>, Aubrey E Thompson<sup>2</sup>, Yan W Asmann<sup>3</sup> and Jean-Pierre A Kocher<sup>1,4\*</sup>

### Abstract

**Background:** Although the costs of next generation sequencing technology have decreased over the past years, there is still a lack of simple-to-use applications, for a comprehensive analysis of RNA sequencing data. There is no one-stop shop for transcriptomic genomics. We have developed MAP-RSeq, a comprehensive computational workflow that can be used for obtaining genomic features from transcriptomic sequencing data, for any genome.

**Results:** For optimization of tools and parameters, MAP-RSeq was validated using both simulated and real datasets. MAP-RSeq workflow consists of six major modules such as alignment of reads, quality assessment of reads, gene expression assessment and exon read counting, identification of expressed single nucleotide variants (SNVs), detection of fusion transcripts, summarization of transcriptomics data and final report. This workflow is available for Human transcriptome analysis and can be easily adapted and used for other genomes. Several clinical and research projects at the Mayo Clinic have applied the MAP-RSeq workflow for RNA-Seq studies. The results from MAP-RSeq have thus far enabled clinicians and researchers to understand the transcriptomic landscape of diseases for better diagnosis and treatment of patients.

**Conclusions:** Our software provides gene counts, exon counts, fusion candidates, expressed single nucleotide variants, mapping statistics, visualizations, and a detailed research data report for RNA-Seq. The workflow can be executed on a standalone virtual machine or on a parallel Sun Grid Engine cluster. The software can be downloaded from <http://bioinformaticstools.mayo.edu/research/maprseq/>.

**Keywords:** Transcriptomic sequencing, RNA-Seq, Bioinformatics workflow, Gene expression, Exon counts, Fusion transcripts, Expressed single nucleotide variants, RNA-Seq reports

**Figure 27:** Snapshot of publication for MAP-RSeq workflow

**From:** Editorial Office [mailto:editorialoffice@oncotarget.com]  
**Sent:** Wednesday, May 03, 2017 2:35 AM  
**To:** Nair, Asha A., M.S.  
**Subject:** Re: New message from contact form

You may use the paper as needed, as long as the source is cited, in accordance with the license we use for all our papers: <http://creativecommons.org/licenses/by/3.0/>


On Sun, Apr 30, 2017 at 10:05 AM, <[info@oncotarget.com](mailto:info@oncotarget.com)> wrote:  
You have new message from contact form

=====  
Name: Asha Nair  
Subject: Permission for use of publication in PhD thesis  
Message: Dear Madam/Sir,

My name is Asha Nair and I am the first author of the publication titled "Circular RNAs and their associations with breast cancer subtypes" published in Oncotarget in 2016. I am a PhD student at the University of Minnesota and this work is part of my final thesis. Hence, I kindly request if you could please provide me with a letter authorizing the use of this publication and the permission to reproduce this work in my thesis.

Please contact me via phone or email if you have any questions.  
Thanks much  
Asha

**Figure 28:** Permission from Oncotarget to reproduce circular RNA publication



## Attribution 3.0 Unported (CC BY 3.0)


This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

### You are free to:

**Share** — copy and redistribute the material in any medium or format


**Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.



---

### Under the following terms:



**Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

**No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

---

### Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

**Figure 29:** Creative Commons license agreement used by Oncotarget

## Circular RNAs and their associations with breast cancer subtypes

Asha A. Nair<sup>1</sup>, Nifang Niu<sup>2</sup>, Xiaojia Tang<sup>1</sup>, Kevin J. Thompson<sup>1</sup>, Liewei Wang<sup>3</sup>, Jean-Pierre Kocher<sup>1</sup>, Subbaya Subramanian<sup>4</sup>, Krishna R. Kalari<sup>1</sup>

<sup>1</sup>Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

<sup>2</sup>Division of Genomic and Molecular Pathology, University of Chicago, Chicago, IL, USA

<sup>3</sup>Department of Pharmacology, Mayo Clinic, Rochester, MN, USA

<sup>4</sup>Division of Basic and Translational Research, University of Minnesota, Minneapolis, MN, USA

**Correspondence to:** Krishna R. Kalari, **email:** Kalari.Krishna@mayo.edu  
Subbaya Subramanian, **email:** subree@umn.edu

**Keywords:** *circular RNA, circ-seq, breast cancer, molecular subtypes, proliferation*

**Received:** February 23, 2016

**Accepted:** October 29, 2016

**Published:** November 05, 2016

### ABSTRACT

Circular RNAs (circRNAs) are highly stable forms of non-coding RNAs with diverse biological functions. They are implicated in modulation of gene expression thus affecting various cellular and disease processes. Based on existing bioinformatics approaches, we developed a comprehensive workflow called Circ-Seq to identify and report expressed circRNAs. Circ-Seq also provides informative genomic annotation along circRNA fused junctions thus allowing prioritization of circRNA candidates. We applied Circ-Seq first to RNA-sequence data from breast cancer cell lines and validated one of the large circRNAs identified. Circ-Seq was then applied to a larger cohort of breast cancer samples ( $n = 885$ ) provided by The Cancer Genome Atlas (TCGA), including tumors and normal-adjacent tissue samples. Notably, circRNA results reveal that normal-adjacent tissues in estrogen receptor positive (ER+) subtype have relatively higher numbers of circRNAs than tumor samples in TCGA. Similar phenomenon of high circRNA numbers were observed in normal breast-mammmary tissues from the Genotype-Tissue Expression (GTEx) project. Finally, we observed that number of circRNAs in normal-adjacent samples of ER+ subtype is inversely correlated to the risk-of-relapse proliferation (ROR-P) score for proliferating genes, suggesting that circRNA frequency may be a marker for cell proliferation in breast cancer. The Circ-Seq workflow will function for both single and multi-threaded compute environments. We believe that Circ-Seq will be a valuable tool to identify circRNAs useful in the diagnosis and treatment of other cancers and complex diseases.

**Figure 30:** Snapshot of publication for Circ-Seq workflow and circular RNA associations found in breast cancer