ELUCIDATION OF PRE-MICRORNA PROFILES OF BREAST CANCERS FOR PATHOGENESIS AND PROGNOSTIC SIGNIFICANCE

WU SEN

UNIVERSITI SAINS MALAYSIA

2025

ELUCIDATION OF PRE-MICRORNA PROFILES OF BREAST CANCERS FOR PATHOGENESIS AND PROGNOSTIC SIGNIFICANCE

by

WU SEN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

September 2025

ACKNOWLEDGEMENT

Upon the occasion of my graduation, I would like to extend my heartfelt gratitude to several individuals whose support has been instrumental in my academic journey. Foremost, I am deeply thankful to my supervisor Dr. Ch'ng Ewe Seng, whose guidance in project design, results discussion, and thesis writing has significantly contributed to my academic progress. Additionally, I am grateful to my co-supervisor Prof. Dr. Tang Thean Hock, who not only provided essential academic guidance but also offered substantial personal support. His assistance in securing a semester's tuition and living expenses was crucial during my most challenging times. I also wish to acknowledge Dr. Citartan and Dr. Siti Aminah for their invaluable help in facilitating my smooth adaptation to the new environment and for their guidance during the experimental phase of my research. I would like to express my sincere gratitude to Prof. Soo-Hwang Teo and Dr. Jia-Wern Pan from Cancer Research Malaysia for providing me access to the raw total RNA-seq data. I am also thankful to Dr. Carsten Raabe for his valuable suggestions, which, along with those from Prof. Teo and Dr. Pan, greatly enhanced the quality of this work. Furthermore, I extend my thanks to the staff of IPS and IPPT for their timely assistance whenever I encountered difficulties. Lastly, I am profoundly appreciative of my parents and family, whose unwavering support has been the cornerstone of my ability to study abroad with peace of mind.

In addition, I would also like to thank the following people whose financial support helped me complete my studies.

Mr. Yeap Hee Kong, Prof. Rabinderjeet, Mr. Saw Teck Chee, Mr. Yew Yew Sun, Mr. Tan Chien Chtze, Mr. Owi Boon Teck, Mdm. See Teik Lan, Mr. Low Sor Geok, and Mdm. Ang Peik Kim.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Ablim1	Actin binding LIM protein 1
ACACB	Acetyl-coa carboxylase beta
ADAM12	ADAM metallopeptidase domain 12
ADIPOQ	Adiponectin, C1Q and collagen domain containing
ADR	Adriamycin
AP1S1	Adaptor related protein complex 1 subunit sigma 1
ATG1	Autophagy-related 1
ATG4D	Autophagy related 4D cysteine peptidase
ATG7	Autophagy related 7
ATM	ATM serine/threonine kinase
AUC	Area under curve
B3GNT5	Beta-1,3-N-acetylglucosaminyltransferase 5
BCL2	BCL2 apoptosis regulator
BECN1	Beclin 1
BLAST	Basic local alignment search tool
BRCA1	BRCA1 DNA repair associated
BRCA2	BRCA2 DNA repair associated
CACNG4	Calcium voltage-gated channel auxiliary subunit gamma 4
CASQ1	Calsequestrin 1
CASQ2	Calsequestrin 2
CCL25	C-C motif chemokine ligand 25
CCND1	Cyclin D1
CD44	CD44 molecule
CDK6	Cyclin dependent kinase 6
CLDN1	Claudin 1
CLDN9	Claudin 9
CLL	Chronic lymphocytic leukemia
CNTFR	Ciliary neurotrophic factor receptor
COL11A2	Collagen type XI alpha 2 chain
COL22A1	Collagen type XXII alpha 1 chain
COL2A1	Collagen type II alpha 1 chain

COL4A4	Collagen type IV alpha 4 chain
COL9A1	Collagen type IX alpha 1 chain
COQ4	Coenzyme Q4
CREB5	CAMP responsive element binding protein 5
CSF1	Colony stimulating factor 1
CXCL16	C-X-C motif chemokine ligand 16
DEG	Differentially expressed gene
DEPM	Differentially expressed pre-microRNA
DLG4	Discs large MAGUK scaffold protein 4
DNMT3B	DNA methyltransferase 3 beta
DOCK3	Dedicator of cytokinesis 3
DOX	Doxorubicin
DPF3	Double PHD fingers 3
ECM	Extracellular matrix
EFNA4	Ephrin A4
EFNA5	Ephrin A5
EGF	Epidermal growth factor
ELOVL4	ELOVL fatty acid elongase 4
EMT	Epithelial-mesenchymal transition
ENY2	ENY2 transcription and export complex 2 subunit
EPS	Exopolysaccharide
eQTL	Expression quantitative trait loci
ER	Estrogen receptor
FADS2	Fatty acid desaturase 2
FAM3D	FAM3 metabolism regulating signaling molecule D
FASLG	Fas ligand
FBXL16	F-box and leucine rich repeat protein 16
FDR	False discovery rate
FGF14	Fibroblast growth factor 14
FOXC1	Forkhead box C1
FOXO1	Forkhead box O1
FSCN1	Fascin actin-bundling protein 1
FUNDC1	FUN14 domain containing 1
G6PC3	Glucose-6-phosphatase catalytic subunit 3

GNAT1	G protein subunit alpha transducin 1
GO	Gene ontology
GPM6B	Glycoprotein M6B
GPR37	G protein-coupled receptor 37
GRM6	Glutamate receptor, metabotropic 6
GSG1L	GSG1 like
GWAS	Genome-wide association studies
HCN1	Hyperpolarization activated cyclic nucleotide gated potassium channel 1
HCN2	Hyperpolarization activated cyclic nucleotide gated potassium and sodium channel 2
HER2	Human epidermal growth factor receptor 2
HER2BC	Human epidermal growth factor receptor 2-positive breast cancer
HGF	Hepatocyte growth factor
HMGB1	High-mobility group box 1
HOXD10	Homeobox D10
IL2RG	Interleukin 2 receptor, gamma chain
IL32	Interleukin 32
IL6	Interleukin 6
ISL	Isoliquiritigenin
KCNC1	Potassium voltage-gated channel subfamily C member 1
KCNK6	Potassium two pore domain channel subfamily K member 6
KCNQ	Potassium voltage-gated channel subfamily Q member
KEGG	Kyoto encyclopedia of genes and genomes
KM	Kaplan-meier
KRAS	KRAS proto-oncogene
LASSO	Least absolute shrinkage and selection operator
LBC	Luminal breast cancer
LC3	Microtubule associated protein 1 light chain 3 alpha
LIN28B	Lin-28 homolog B
M7G	7-methylguanosine
MAPT	Microtubule associated protein tau
METTL3	Methyltransferase 3
miEAA	Mirna Enrichment Analysis and Annotation Tool
MIENTURNET	Microrna enrichment turned networks
miR-seq	Microrna-seq/microrna sequencing

MR	Mendelian randomization
mTOR	Mechanistic target of rapamycin kinase
mTORC1	Mechanistic target of rapamycin kinase complex 1
MYBL2	MYB proto-oncogene like 2
MyBrCa	Malaysian breast cancer genetic study
MYCN	MYCN proto-oncogene
MZF1	Myeloid zinc finger 1
NCOA7	Nuclear receptor coactivator 7
NGS	Next-Generation sequencing
NLRC5	NLR family CARD domain containing 5
NSCLC	Tumor protein p53
nts	Nucleotides
NTSR1	Neurotensin receptor 1
P27	P27 protein
PARP	Poly (ADP-ribose) polymerase
PCA	Principal component analysis
PDCD4	Programmed cell death 4
PI3K	Phosphatidylinositol 3-kinase
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PPI	Protein-protein interaction
PR	Progesterone receptor
Pre-microRNA	Precursor microRNA
PTEN	Phosphatase and tensin homolog
RAB5A	Rab5a
RECK	Reversion inducing cysteine rich protein with kazal motifs
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
RPL8	Ribosomal protein L8
SCUBE2	Signal peptide, CUB domain and EGF like domain containing 2
SNP	Single nucleotide polymorphism
STAT1	Signal transducer and activator of transcription 1
STC1	Stanniocalcin 1
STMN1	Stathmin 1
TCGA	The cancer genome atlas

TIMP3	TIMP metallopeptidase inhibitor 3
TMM	Trimmed Mean of M-values
TNBC	Triple negative breast cancer
TP53	Tumor protein p53
TPM	Transcripts-per-million
ULK1	Unc-51 like autophagy activating kinase 1
UTR	Untranslated region
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
ZEB1	Zinc finger E-box binding homeobox 1
ZEB2	Zinc finger E-box binding homeobox 2

LIST OF APPENDICES

PENJELASAN PROFIL PRA-MIKRORNA KANSER PAYUDARA UNTUK TUJUAN PATOGENESIS DAN KEPENTINGAN PROGNOSIS

ABSTRAK

Kanser payudara diklasifikasikan kepada kanser payudara luminal positif reseptor hormon (LBC), kanser payudara positif reseptor faktor pertumbuhan epidermis manusia 2, dan kanser payudara triple-negatif (TNBC). Mikronya prekursor (pre-mikroRNA), yang biasanya membentuk struktur pin rambut dengan panjang antara 65 hingga 80 bes, telah ditunjukkan memainkan peranan penting dalam karsinogenesis kanser payudara. Ia boleh diandaikan bahawa pre-mikroRNA boleh dikenalpasti melalui penjujukan RNA keseluruhan (RNA-seq). Untuk menguji hipotesis ini, algoritma novel bernama b-mer telah direka dan disahkan membangunkan set data prjna749047. Keputusan menunjukkan bahawa b-mer berjalan dengan cekap tetapi juga memberikan anotasi pra-mikro-mikro yang unggul berbanding dengan kaedah tradisional "pemetaan ke genom rujukan". 907 sampel kanser payudara daripada set data MyBrCa telah diprofil menggunakan b-mer dan perbandingan dibuat antara profil pra-micrornas dan profil microrna matang yang diperoleh daripada set data atlas genom kanser (TCGA). Sepuluh pra-mikrorna yang dinyatakan secara berbeza telah dikenal pasti secara biasa dalam kedua-dua mybrca dan tcga. Empat tandatangan pra-micrornas telah dibina dan tiga gen sasaran (MYBL2, FAM3D, B3GNT5) telah disahkan sebagai berkaitan sebab akibat dengan kanser payudara dalam kedua-dua kumpulan Jepun dan Eropah. Secara kesimpulannya, b-mer memudahkan pemprofilan pra-mikrorna secara langsung daripada jumlah data rna-seq mentah kanser payudara, dan mengenal pasti senarai gen dan pra-mikrorna yang penting untuk perkembangan dan prognosis penyakit. Penemuan kajian ini

menyerlahkan peranan kritikal pra-mikrorna dalam karsinogenesis kanser payudara dan potensinya sebagai faktor prognostik.

ELUCIDATION OF PRE-MICRORNA PROFILES OF BREAST CANCERS FOR PATHOGENESIS AND PROGNOSTIC SIGNIFICANCE

ABSTRACT

Breast cancer is classified into hormone receptor-positive luminal breast cancer (LBC), human epidermal growth factor receptor 2-positive breast cancer, and triple-negative breast cancer (TNBC). Precursor microRNAs (pre-microRNAs), typically form hairpin structures with a length from 65 to 80 bases, are shown to play crucial roles in breast cancer carcinogenesis. It can be hypothesized that pre-microRNAs could be identified through total RNA sequencing (RNA-seq). To test this hypothesis, a novel algorithm named B-mer was designed and validated using the dataset PRJNA749047. The results demonstrated that B-mer runs efficiently but also provides superior annotation of pre-microRNAs compared to the traditional "Mapping-to-reference-genome" method. 907 breast cancer samples from MyBrCa dataset were profiled using B-mer and comparisons were made between pre-microRNAs profiles and mature microRNA profiles obtained from The Cancer Genome Atlas (TCGA) dataset. Ten differentially expressed pre-microRNAs were identified commonly in both MyBrCa and TCGA. A four pre-microRNAs signature was constructed and three target genes (MYBL2, FAM3D, B3GNT5) were validated as causally related to breast cancer in both Japanese and European groups. In conclusion, B-mer facilitates the profiling of pre-microRNAs directly from raw total RNA-seq data of breast cancer, and identifies a list of genes and pre-microRNAs that are pivotal for the development and prognosis of the disease. The findings of this study highlight the critical role of pre-microRNAs in breast cancer carcinogenesis and their potential as prognostic factors.

CHAPTER 1

INTRODUCTION

1.1 Background

Breast cancer is the common prevailing cancer affecting women worldwide and represents an important public health challenge (Sung et al., 2021). The incidence of breast cancer has been increasing globally. According to the Malaysia National Cancer Registry report published in 2024, age-standardized incidence-ASR of breast cancer has seen an important rise from 34.1 per 100,000 among the period of 2012 to 2016 to 38.9 per 100,000 during 2017 and 2021 ("Malaysian National Cancer Registry Report (NCR) 2017-2021," 2024). This up trend is brooding of broad patterns observed in Asian countries as reported by the World Health Organization (Kim et al., 2025). The rise in incidence rates underscores the need for enhanced screening, detection, and curative strategies to combat the growing burden of breast cancer breast cancer.

Breast cancer is basically divided into three subtypes based on the hormone receptors that are present or absent and the expression of human epidermal growth factor receptor 2 (HER2) protein: luminal breast cancer (LBC), HER2-overexpression breast cancer (HER2BC), and triple-negative breast cancer (TNBC) (Jin et al., 2023). Each subtype manifests unique molecular features and clinical behaviours, which influences treatment decisions and outcomes (Lehmann et al., 2011; Romero-Cordoba et al., 2020). Luminal breast cancer (LBC) is characterized by the presence of estrogen receptor (ER) and/or progesterone receptor (PR), accounts for the majority of breast cancer cases. Generally, these cancers have a relatively good prognosis and respond well to hormonal therapies (Jin et al., 2023). Tamoxifen and aromatase inhibitors, for

example, block the effects of estrogen on cancer cell growth (Sørlie et al., 2001). HER2-overexpressed breast cancer (HER2BC) is marked by the overexpression of the human epidermal growth factor receptor 2 (HER2) protein. HER2BC tends to be more aggressive but can respond to targeted therapies, such as trastuzumab and pertuzumab, which specifically inhibit the HER2 protein and its downstream signalling pathways (Baselga & Swain, 2009). Triple-Negative Breast Cancer (TNBC) is known for lacking of ER, PR, and HER2 receptors, making it the most challenging subtype to treat. It is often associated with poorer outcomes and requires more aggressive treatments, such as chemotherapy, due to the absence of targeted therapies (Lehmann et al., 2011).

Molecular heterogeneity and wide-ranging clinical presentations of breast cancer are led by its divers genetic, epigenetic, environmental factors, which contribute to the development and progression (Ali et al., 2014). The molecular heterogeneity of breast cancer underscores the complex mechanism of disease initiation and necessitates personalized treatment approaches to improve patient outcomes and reduce mortality rates (Pasha & Turner, 2021).

MicroRNAs (microRNAs), a kind of small non-coding RNAs with typically 18-25 nucleotides in length, come from a multi-step process of transcription (Rogers & Chen, 2013). The biogenesis of microRNAs begins in the nucleus with the transcription of primary microRNAs (pri-microRNAs) by RNA polymerase II. These long transcripts, which can extend up to thousands of bases, are processed by the Drosha-DGCR8 complex to generate precursor microRNAs (pre-microRNAs) (Lee et al., 2004). Pre-microRNAs, typically 65-80 bases long and forming hairpin loop structures, are exported to the cytoplasm by Exportin-5. In the cytoplasm, pre-

microRNAs are processed into mature microRNA duplexes with the assistance of another RNase III enzyme, Dicer{Su, 2025 #1577}. One strand of the duplex, known as the guide strand, is incorporated into the RNA-induced silencing complex (RISC), where it guides the complex to target mRNAs based on sequence complementarity, leading to mRNA degradation or translational repression (Lund et al., 2006). The lifespan of pre-microRNAs in cells is however unknown and generally regarded as transient intermediaries for mature microRNAs.

MicroRNAs have emerged as important regulators of gene expression in breast cancer and other diseases for regulation of genes' expression post-transcriptionally by binding to the un-translational region (UTR) of target mRNAs, leading to their degradation or inhibition of transcription (Ambros, 2004). MicroRNAs control various cellular processes, including proliferation, differentiation, and apoptosis, making them pivotal in cancer biology (Rupaimoole & Slack, 2017). MicroRNAs can function as oncogenes (oncomiRs) or tumor suppressors, depending on their target genes. For instance, overexpression of certain microRNAs can lead to the downregulation of tumor suppressor genes, promoting cancer cell proliferation and survival (Johnson et al., 2007; Pan et al., 2013). Conversely, the loss of tumor-suppressive microRNAs can result in the upregulation of oncogenes, contributing to cancer progression and metastasis.

With the development of high throughput sequencing technology, microRNA-seq (miR-seq) or microRNA microarray technologies are typically used to evaluate microRNA profiles, which basically focus on the mature microRNAs (Eipper-Mains et al., 2011). Current microRNA profiling techniques have provided valuable insights into the expression patterns of mature microRNAs in health and disease. However,

several issues are often ignored by these analyses. Firstly, there is a disparity in the measurement methods for mRNA and microRNAs within the same studies, resulting in a gap when studying the relation between the mRNA and microRNA data (Faridani et al., 2016). Consequently, conducting a comparable comparison of mRNAs and microRNAs to construct a precise regulation network within a single study becomes unattainable. Second, it is essential to acknowledge that some mature microRNAs undergo rapid degradation once released from RNA-induced silencing complex (RISC) or stress response started (Chatterjee & Großhans, 2009; Leung & Sharp, 2010; Smibert et al., 2013). For example, stress leading to a DNA damage will decrease the level of microRNA-125b, while the mechanisms behind this are still unclear. Consequently, both miR-seq and microarray techniques are incapable of accurately capturing the actual expression levels of microRNAs, leading to potential distortions in the results (Jayaprakash et al., 2011). Moreover, previously published microRNA-seq outcomes, lack the inclusion of new microRNAs due to database limitations, with the annotation of microRNAs often lagging and remaining ambiguous (Kozomara et al., 2019). Taking the microRNA expression of breast cancer released by TCGA as an example, there is no more update following the new finding of microRNAs. Consequently, the process of updating the microRNA profile through a complete re-run of microRNA-seq analysis is time-consuming. Lastly, scrutiny of public databases for microRNA profiles reveals a scarcity of data pertaining to Asian/African samples, introducing biases in the signatures derived from microRNAs (Popejoy & Fullerton, 2016; Spratt et al., 2016).

Total RNA sequencing (RNA-seq) technology, which captures all RNA in a sample, offers an opportunity to study pre-microRNAs. However, the lack of a specific algorithm to extract pre-microRNAs from total RNA-seq data has limited its

application in microRNA research. Developing such an algorithm could provide a more comprehensive view of microRNA regulation and uncover novel biomarkers for breast cancer.

1.2 Problem Statement

The central problem of this study can be summarized as follows: Is it possible to develop a robust algorithm to annotate and quantify pre-microRNAs from total RNA-seq data, and how can pre-microRNA profiles enhance the understanding of breast cancer and improve prognostic capabilities?

This problem addresses four key issues:

Algorithm Development: Is it possible to design a reliable computational method to identify and quantify pre-microRNAs from total RNA-seq raw reads?

Data Integrity and Validation: How do the pre-microRNA profiles generated by this algorithm compare with mature microRNA profiles obtained from databases like The Cancer Genome Atlas (TCGA)?

Functional Insights: What additional regulatory information can pre-microRNAs provide about breast cancer might have been overlooked in previous studies focusing on mature microRNAs?

Prognostic Value: Can pre-microRNA profiles be used to develop a prognostic signature for breast cancer?

1.3 General Objectives

The general objective of this study is to develop a robust algorithm capable of annotating and quantifying precursor microRNAs (pre-microRNAs) from total RNA sequencing (RNA-seq) data, then leverage these pre-microRNA profiles to gain deeper insights into the molecular mechanisms of breast cancer and to provide novel diagnostic and prognostic tools.

1.3.1 Objective 1: To Develop and Validate a Pre-microRNA Annotation and Quantification Algorithm

Firstly, the study aims to design and develop a computational algorithm that accurately annotates and quantifies pre-microRNAs directly from raw total RNA-seq reads and to validate the performance of algorithm by comparing with those derived from mature microRNA sequencing (miR-seq) and total RNA-seq data.

1.3.2 Objective 2: To Profile Pre-microRNAs in Breast Cancer and Uncover Functional Insights

Secondly, this study aims to generate pre-microRNA expression profiles for breast cancer cohorts, specifically those curated by the Malaysian Breast Cancer Genetic Study (MyBrCa). The study conducts comprehensive comparative analyses based on pre-microRNA profiles of Luminal Breast Cancer (LBC) and Triple-Negative Breast Cancer (TNBC) to:

•Uncover subtype-specific and cancer-specific pre-microRNAs with potential biological significance

• Predict target genes of significant pre-microRNAs using established bioinformatic tools like miRBase and TargetScan

•Perform pathway enrichment analyses to understand the biological processes and pathways regulated by these pre-microRNAs

• Construct microRNA-gene interaction networks to visualize regulatory relationships and identify key regulatory hubs

1.3.3 Objective 3: To Evaluate the Prognostic Value and Causal Relationship of Pre-microRNAs in Breast Cancer

Lastly, this study aims to evaluate the prognostic value of pre-microRNA profiles in breast cancer. Additionally, the study tests for causal relationships between the predicted target genes of pre-microRNAs and breast cancer development or progression.

1.4 Study Design

This study leverages advanced bioinformatics techniques and extensive total RNA sequencing datasets to develop and validate a novel algorithm for annotating and quantifying precursor microRNAs (pre-microRNAs) and then applies the algorithm to elucidation the importance of pre-microRNA species in breast cancer. The research is structured into several key phases, each focusing on distinct aspects of algorithm development, data analysis, and the application of the algorithm for pre-microRNA profiling in breast cancer.

1. Algorithm development and validation: The first phase involves designing and implementing a computational algorithm to accurately annotate and quantify pre-microRNAs from total RNA-seq raw reads. This algorithm integrates genomic annotations and computational methods to distinguish pre-microRNAs from other RNA species (in **Chapter 3**).

2. Generation of pre-microRNA expression profiles for breast cancer samples collected by MyBrCa project by the developed algorithm and comprehensive analysis of pre-microRNA profile (in **Chapter 4**).
3. Evaluation of the importance of pre-microRNAs obtained by comparison between LBC and TNBC with both mendelian randomization analysis of target genes of differentially expressed pre-microRNAs and prognostic signature construction of pre-microRNAs (in **Chapter 5 and Chapter 6**).

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

2.1.1 Background on Breast Cancer

Breast cancer, the most common cancer affecting women worldwide, is a molecularly heterogeneous disease (Perou et al., 2000). This heterogeneity implies that breast cancer encompasses a variety of subtypes, each with distinct molecular and clinical characteristics. The classification of breast cancer into its various subtypes is crucial for determining the most appropriate treatment strategies and predicting patient outcomes (Goldhirsch et al., 2011).

One of the most popular methods to determine the subtype of breast cancer is molecular tests for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (Hammond et al., 2010). These molecular markers are pivotal in guiding the diagnosis and management of breast cancer. Based on the results of these molecular tests, breast cancer can be classified into three main subtypes: luminal breast cancer (LBC), HER2-overexpressed breast cancer (HER2BC), and triple-negative breast cancer (TNBC).

Luminal Breast Cancer (LBC) is characterized by high expression of estrogen receptor (ER) and/or progesterone receptor (PR), and low expression of HER2 (Koboldt et al., 2012). Luminal breast cancers are further divided into luminal A and luminal B subtypes, based on their proliferation rates and additional molecular characteristics (Goldhirsch et al., 2011). Luminal A tumors generally have a better prognosis and respond well to hormonal therapies, such as tamoxifen and aromatase inhibitors, which target the ER pathway. Luminal B tumors, although still hormone

receptor-positive, tend to be more aggressive than luminal A tumors and may require additional treatments, such as chemotherapy (Goldhirsch et al., 2013; Sørlie et al., 2001).

HER2-Overexpressed Breast Cancer (HER2BC) is characterized by the overexpression of the HER2 protein, a receptor tyrosine kinase involved in cell growth and differentiation(Slamon et al., 1987). HER2BC tends to be more aggressive than luminal subtypes but can be effectively treated with targeted therapies, such as trastuzumab (Herceptin) and pertuzumab (Perjeta), which specifically inhibit the HER2 protein (Baselga et al., 2012; Slamon et al., 2001). These targeted treatments have significantly improved the prognosis for patients with HER2BC breast cancer.

Triple-Negative Breast Cancer (TNBC) is defined by the absence of ER, PR, and HER2 expression. This subtype accounts for about 15-20% of all breast cancers and is known for its aggressive behaviour and poor prognosis(Foulkes et al., 2010). TNBC is more likely to metastasize early and lacks targeted therapies, making treatment more challenging. Chemotherapy remains the mainstay of treatment for TNBC, although ongoing research is exploring new therapeutic options, such as immunotherapy and PARP inhibitors (Tutt et al., 2010).

In comparison to LBC, HER2BC and TNBC subtypes are characterized by a more aggressive biology, resulting in a higher level of difficulty in treatment and a poorer prognosis(Johnson et al., 2020). These aggressive subtypes are associated with higher rates of recurrence and metastasis, underscoring the need for effective treatment strategies and continuous research to improve outcomes.

The heterogeneity of breast cancer emphasizes the importance of personalized medicine. Understanding the molecular profile of each tumor allows clinicians to tailor

treatments to individual patients, optimizing therapeutic efficacy and minimizing unnecessary side effects (Perou et al., 2000). Advances in molecular biology and genomics continue to enhance the ability to classify breast cancer subtypes more precisely, paving the way for novel treatment approaches and improved patient care (Curtis et al., 2012).

2.1.2 Overview of microRNAs

MicroRNA is a collection of short non-coding RNAs with lengths from 18-25 nts. The biogenesis of microRNAs involves multiple steps and cellular compartments. The process begins in the nucleus and concludes in the cytoplasm, involving both canonical and non-canonical pathways (Bartel, 2004). Primary microRNAs (Pri-microRNAs) are firstly transcribed in the nucleus intermediated by RNA polymerase II, consisting of multi hairpin structures (Lee et al., 2004). Each hairpin structure in pri-microRNAs is cleaved to produce precursor microRNA (pre-microRNA). Exportin-5 exports pre-microRNAs to cytoplasm, which allows per-microRNA to be cleaved into double-stranded microRNA complex, which consists one or two mature microRNAs (Kim, 2004).

MicroRNAs are classified into two subtypes based on their location relative to coding genes: intronic microRNAs and intergenic microRNAs. Intronic microRNAs are situated within the introns of host genes and their transcription depends on these host genes. In contrast, intergenic microRNAs are located between protein-coding genes and exhibit more independent expression patterns. According to the latest human genome reference and miRBase v22 annotations, 2882 microRNAs are distributed across all chromosomes including the sex chromosomes (XY). Chromosome 19 contains the highest number of microRNAs, with 242, followed by Chromosome 1 and the X chromosome, which have 233 and 193 microRNAs, respectively. This

distribution suggests that some microRNAs may play roles in regulating gender-specific diseases. Notably, Chromosome Y has only 4 microRNAs. When considering chromosome sizes, the top five chromosomes with the highest microRNA density are Chromosomes 19, 17, 14, 22, 16, and X. (see Figure 2.1).

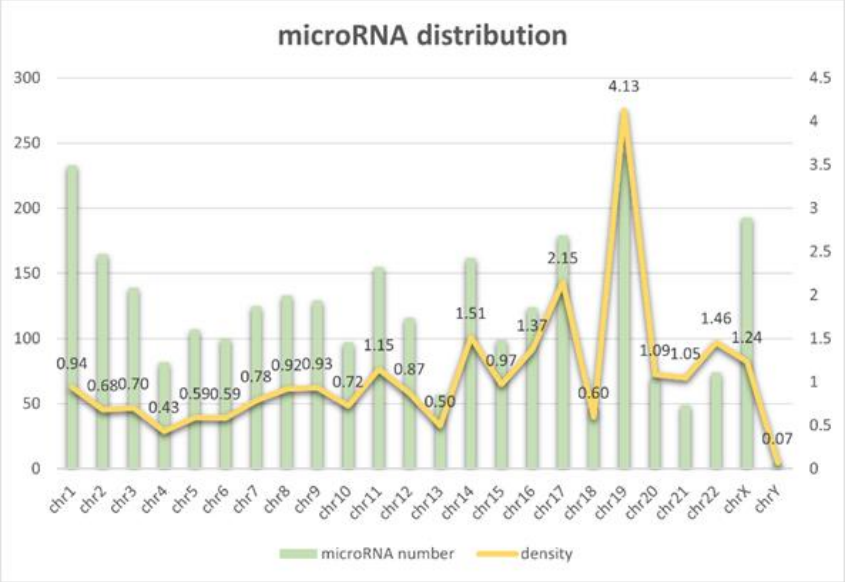

Figure 2.1 Distribution of microRNAs.

The figure shows the number (green bars) and density (yellow line) of microRNAs across different chromosomes. The green bars represent the total number of microRNAs on each chromosome, while the yellow line indicates the microRNA density, calculated as the number of microRNAs per million bases. Chromosome 19 (chr19) exhibits a significant peak in both microRNA number and density, suggesting a higher concentration of microRNAs on this chromosome. The numbers at the bottom of each bar and along the yellow line denote the specific microRNA count and density for each chromosome, respectively. This figure was summarized from the annotation of the reference human genome from NCBI (O'Leary et al., 2024).

Function of microRNAs are always thought to be negative to target genes for their imperfect complementarity binding to the un-translational region (UTR) of target genes (Valencia-Sanchez et al., 2006). MicroRNAs are not only essential in regulating the timing and patterns of development, but also associated to diseases, especially for cancers (Lan et al., 2015).

2.2 Function of MicroRNAs in Breast Cancer

MicroRNAs play a crucial role in regulating the expression of their target genes, thereby influencing various cellular functions (Garzon et al., 2017). In the context of breast cancer, 656 microRNAs and their interactions have been identified as either promoting or inhibiting therapeutic development. Approximately 10% of these interactions have been experimentally validated, demonstrating their involvement in tumorigenesis, development, apoptosis, and autophagy of breast cancer cells. For instance, miR-181 regulates the expression of serine-protein kinase ATM (ATM) during the initiation stage of cancer (Wang et al., 2011). Additionally, miR-373 stabilizes the breast cancer microenvironment by interacting with CD44 molecule (Iczkowski, 2010; Monchusi & Kaur, 2022). Furthermore, miR-222 and miR-221 target cyclin-dependent kinase inhibitor 1B (p27), influencing cell proliferation and contributing to therapy resistance (Wang et al., 2016). The Figure 2.2 below shows how microRNA regulate the autophagy process as an example.

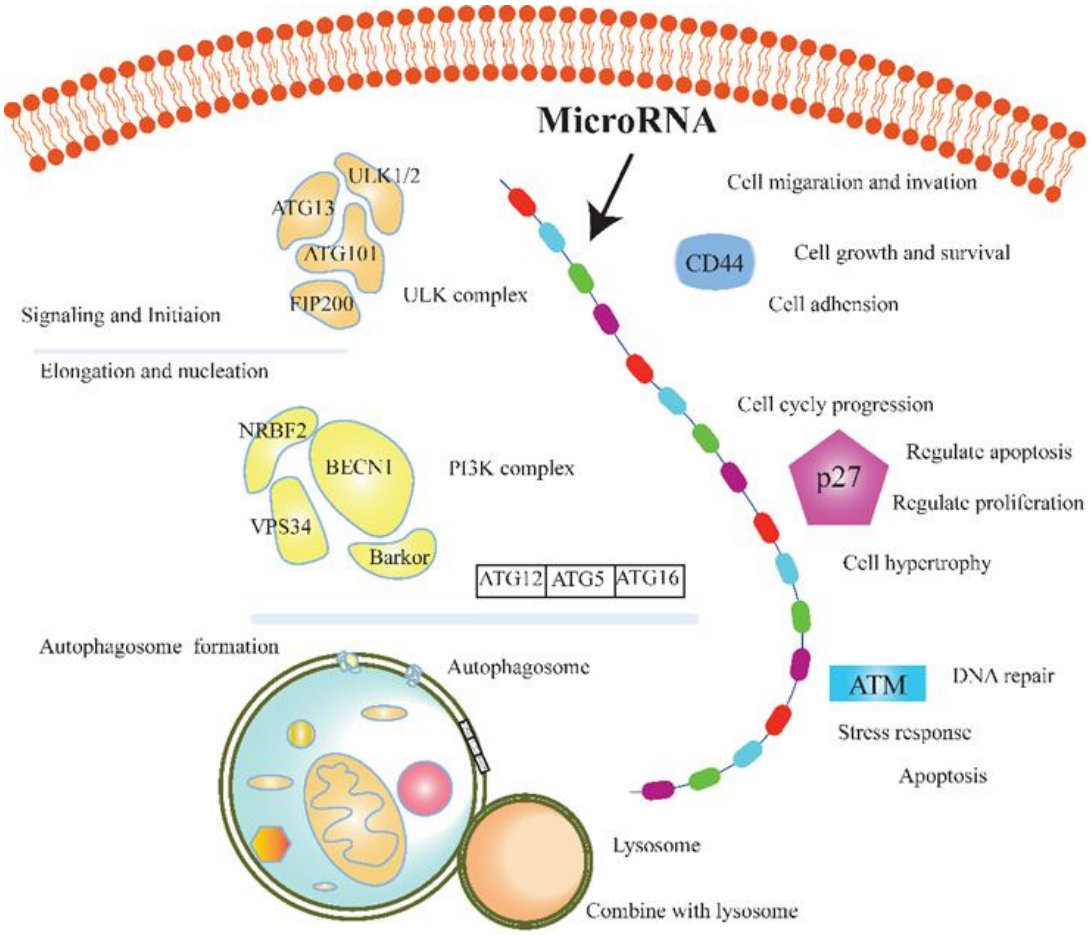

Figure 2.2 MicroRNAs regulate autophagy genes in breast cancer.

The right part of microRNA shows selected genes regulated by microRNAs and their functions in breast cancer. CD44 is related to cell growth and migration which will assist cancer cells to survival from immune reaction or treatment. P27 regulates the cell cycle by different pathways. And ATM is related to DNA repair and stress response. The left part shows the major steps of autophagy and pathways involved in autophagy progress.

2.2.1 MicroRNAs as tumor suppressors

Tumor suppressor microRNAs are particularly important because they counteract oncogenic processes by suppressing genes that drive tumor growth, invasion, and metastasis (Behl et al., 2020). Their loss or downregulation can lead to the activation of oncogenic pathways, contributing to cancer progression and poor prognosis (Bartel, 2009).

Several key microRNAs have been identified as crucial tumor suppressors, such as miR-15a, miR-16, miR-34a, miR-143 and miR-145. The miR-15a/16 cluster and miR-34a are known to target the anti-apoptotic protein BCL2, which is involved in preventing programmed cell death and uncontrolled cell death. Downregulation of miR-15a/16 is frequently observed in various cancers, including chronic lymphocytic leukemia (CLL) and solid tumors like breast and prostate cancer. The loss of miR-15/16 cluster contributes to enhanced cancer cell survival and resistance to therapeutic interventions (Calin et al., 2002; Cimmino et al., 2005). Similarly, miR-34a, which is directly regulated by the tumor suppressor p53, plays a pivotal role in controlling cell cycle progression and apoptosis (Bommer et al., 2007). By targeting genes such as CDK6 and BCL2, miR-34a helps maintain cellular homeostasis and prevent uncontrolled cell growth. Its downregulation is linked to a variety of malignancies, underscoring its role as a crucial tumor suppressor (Ren et al., 2018). Additionally, miR-143/145, which regulates cell growth and differentiation, targets oncogenes like KRAS. Their reduced expression has been associated with colorectal cancer and lung cancer, further highlighting their importance in cancer biology (Tavanafar et al., 2017).

The clinical implications of tumor suppressor microRNAs are profound. They can serve as biomarkers for cancer diagnosis, prognosis, and response to treatment. Alterations in microRNA expression levels often reflect tumor status and can provide valuable prognostic information. For example, specific microRNAs may indicate the presence or aggressiveness of a tumor, aiding in personalized treatment strategies (Sharma et al., 2010). Therapeutically, strategies aimed at restoring the function of downregulated tumor suppressor microRNAs hold significant promise. Approaches such as microRNA mimics, which introduce synthetic microRNAs to restore normal function, and antagomiRs, which inhibit the activity of oncogenic microRNAs, are

being actively investigated (Rupaimoole & Slack, 2017; van Rooij & Olson, 2012). Additionally, gene therapy techniques are being explored to enhance the expression of tumor suppressor microRNAs in cancer cells, offering potential new avenues for effective treatment (Koturbash et al., 2011). The ongoing research into the roles and therapeutic potential of tumor suppressor microRNAs underscores their importance in advancing cancer diagnosis and treatment.

2.2.2 MicroRNAs as oncogenes

A growing number of research has established that certain microRNAs function as oncogenes, or "oncomiRs" in various cancers during the past twenty years. These microRNAs contribute to cancer initiation, progression, and metastasis by promoting cell proliferation, inhibiting apoptosis, and facilitating angiogenesis.

Several microRNAs have been verified with functions to affect cancer initiation in various cancers. For example, miR-21 has been extensively studied for its role as an oncogene in multiple cancer types, including breast, lung, and colorectal cancer. In breast cancer, miR-21 inhibits the expression of STAT3 and promotes the proliferation of breast cancer cells (Zhang et al., 2016). In other cancers, miR-21 promotes tumorigenesis by targeting tumor suppressor genes such as reversion inducing cysteine rich protein with kazal motifs (RECK) and PTEN, leading to enhanced cell survival and proliferation (Wu et al., 2017; Zhang et al., 2008). As an oncogene, miR-155 has also been studied in breast cancer and lung cancer. In breast cancer, miR-155 works with the cooperation with genes related to MAPK pathway, which play crucial roles in the initiation of cancer (Martin et al., 2014). A specific variant of the miR-155 leads to a high rate of cell growth in lung cancer, which indicating the importance of miR-155 (Xie et al., 2015).

The dysregulation of microRNAs as oncogenes can also result in the activation of signaling pathways that are crucial for cancer development. For example, the overexpression of miR-17/92 cluster in cancers such as lymphoma and lung cancer has been linked to the activation of the PI3K/AKT and MYC signaling pathways, which are involved in cell growth, proliferation, and survival (Olive et al., 2009). This underscores the multifaceted role of microRNAs in cancer biology and their potential as therapeutic targets.

Furthermore, microRNAs play a pivotal role in facilitating cancer metastasis. For instance, miR-10b has been implicated in the promotion of metastasis in breast cancer. Ma et al. (Ma et al., 2007) demonstrated that miR-10b initiates metastatic processes by suppressing the expression of homeobox D10 (HOXD10), a gene that inhibits cell migration and invasion. This downregulation allows cancer cells to detach from the primary tumor and spread to distant sites. Another microRNA which plays vital roles in cancer metastasis is miR-21. MiR-21 targets both PTEN and LATS1 and leads to a malignant phenotype in both renal cancer and hepatocellular cancer (An et al., 2017; Bao et al., 2013).

In conclusion, microRNAs act as potent oncogenes in a variety of cancers by regulating genes involved in cell proliferation, apoptosis, and metastasis. Their role in tumor biology highlights their potential as biomarkers for cancer diagnosis and as therapeutic targets. Further research is necessary to fully understand the mechanisms by which microRNAs contribute to oncogenesis and to translate these findings into effective cancer therapies. The Figure 2.3 below lists the connection among tumor process and microRNA/gene axis which were validated in breast cancer (with autophagy as example).

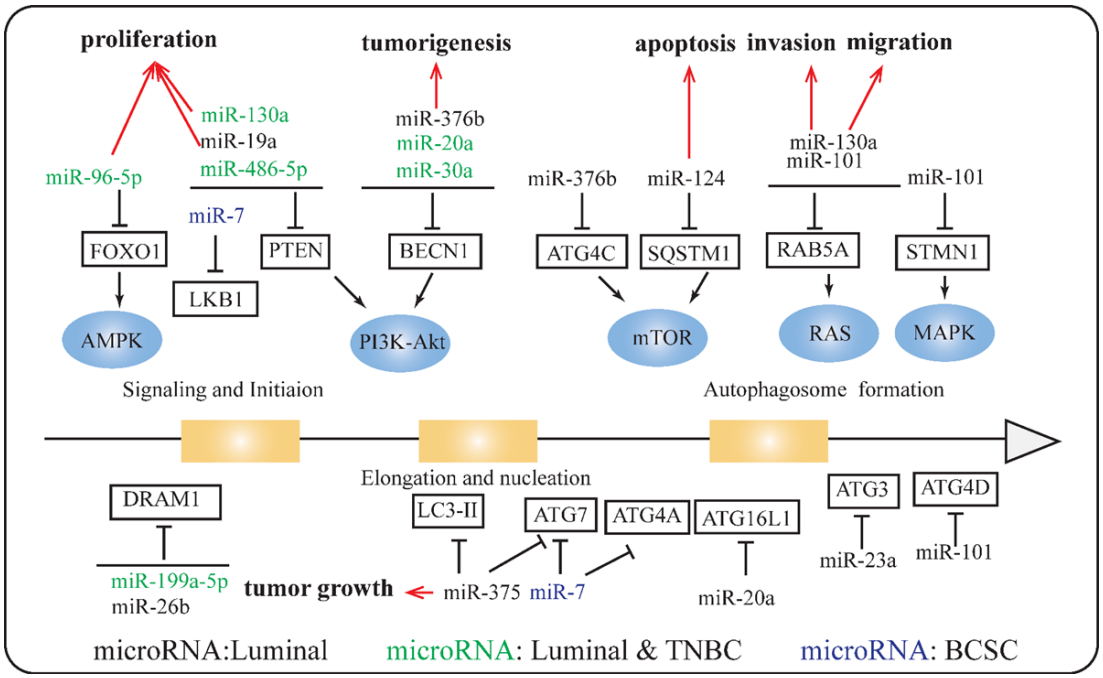

Figure 2.3 Overall direct interactions among tumor process and microRNA/gene axis in different subtypes of breast cancer.

Yellow colored bars represent the major stages of autophagy. MicroRNAs with different colors refer to the subtypes of breast cancer where interactions are verified. Red arrows show how the interactions affect cell destiny.

2.2.3 MicroRNAs related to therapy of Breast cancer

How microRNAs affect the therapeutic efficacy of drugs or radiotherapy in breast cancer has also been investigated. Most of the studies related to the role of microRNA in therapy take luminal subtype as a representative of breast cancer, few cases treat luminal and TNBC cell lines at the same time. However, no study pays attention to all subtypes of breast cancers.

Some of microRNAs have been demonstrated to increase chemosensitivity of luminal cancer cell lines to variable degrees by directly involved autophagic processes. Downregulation of miR-101, for instance, increases the expressions of stathmin 1 (STMN1), RAB5A and autophagy related 4D cysteine peptidase (ATG4D), involving almost all stages of autophagy (Cecconi, 2011). This would sensitize resistant MCF-7

and T47D cancer cell lines to 4-OHT, the main metabolite of tamoxifen. MiR-21 on the other hand increases sensitivity of cancer cells to tamoxifen and fulvestrant by targeting PTEN at the very early stage of autophagic process (Yu et al., 2016). MiR-25 targets unc-51 like autophagy activating kinase 1 (ULK1) in autophagy signaling and initiation step of autophagy, enhancing chemosensitivity of luminal breast cancers, especially, to isoliquiritigenin (ISL), a kind of autophagy inducer(Wang et al., 2014). Generally, ULK1 also known as autophagy-related 1 (ATG1) is expressed under the hypoxia pressure and once it is phosphorylated, FUN14 domain containing 1 (FUNDC1) is produced and it will combine with LC3 to increase the autophagosome formation. And this regulation is suppressed by mTOR signaling pathway when mTORC1 complex is phosphorylated (Kim et al., 2011; Liu et al., 2012; Wu et al., 2014).

Studies have also reported the involvement of microRNAs in relation to other biological process and their effects on commonly employed chemotherapeutic drugs such as Doxorubicin (DOX) and adriamycin (ADR) in luminal breast cancer. For instance, miR-142-3p targets high-mobility group box 1 (HMGB1), leading to increased apoptosis and higher sensitivity to DOX (Liang et al., 2020). MiR-221-3p, which is regulated by methyltransferase 3 (METTL3), enhances the resistance to adriamycin (ADR) by targeting HIPK2 (Pan et al., 2021). Ma et al. (2015) proposed that in HER2+ cells, miR-542-3p may play a partial role in trastuzumab (Herceptin) resistance by targeting genes involved in the PI3K-Akt pathway. However, further investigation is needed to confirm this hypothesis and better elucidate the regulatory network between microRNAs and therapy (Ma et al., 2015).

In summary, microRNAs work in complex network which influence the effect of chemo-therapy to cancers. It requires more attention to the detailed mechanism behind the effect of microRNAs in therapy.

2.3 Sequencing Technology and Its Impact on Breast Cancer Research

2.3.1 Advances in sequencing technologies

The advent of advanced sequencing technologies has significantly shaped the landscape of breast cancer research, offering new insights into the genetic and molecular mechanisms driving the disease. These technologies allow researchers to comprehensively analyze the genomic and transcriptomic alterations in breast cancer, leading to more targeted and personalized therapeutic approaches. Sections below listed key advances in sequencing technologies and their impact on breast cancer research.

2.3.1(a) Next-Generation sequencing (NGS)

Next-Generation Sequencing (NGS) has revolutionized breast cancer research by enabling large-scale genomic analyses at high speed and relatively low cost (Conway & Warner, 2024). NGS technologies, such as those offered by Illumina, provide deep coverage of the genome, facilitating the identification of somatic mutations, copy number variations, and gene fusions specific to breast cancer subtypes (Sheffield, 2022).

NGS has been instrumental in characterizing the genetic heterogeneity of breast cancer, allowing for the identification of mutations in key oncogenes and tumor suppressor genes, such as BRCA1, BRCA2, TP53, and PIK3CA (Condorelli, 2024). These insights have led to the development of targeted therapies, including PARP

inhibitors for BRCA1/2 mutation carriers, and PI3K inhibitors for PIK3CA mutations (Miller & Ledermann, 2016; Pascual & Turner, 2019).

Beyond whole-genome sequencing (WGS) and whole-exome sequencing (WES), NGS has also enabled RNA sequencing (RNA-seq), which provides a comprehensive profile of gene expression (Russo, 2023). RNA-seq has been used to identify aberrant splicing events, gene fusions, and expression changes that contribute to breast cancer pathogenesis (Perou et al., 2000; Riaz et al., 2023). This level of detail allows researchers to identify not only individual mutations but also transcriptomic alterations that can serve as biomarkers for diagnosis, prognosis, and treatment response (Li et al., 2020).

By total RNA-seq sequencing, which contains almost all RNA species in the library excepts for ribosome RNA, small RNA such as microRNA species could be observed in the expression profile (Yang et al., 2024). Annotation of total RNA-seq raw data provides an overview of pre-microRNA species profile. However, profile of pre-microRNA species obtained by traditional tools for RNA-seq analysis differs for various reasons (Fuller et al., 2024).

Firstly, traditional tools for annotation and quantification of genes from raw reads of total RNA-seq will lead to a lag out of the new findings of microRNAs for the requirement of high-quality reference genome, which won't be updated frequently (Gennari, 2021). In such situation, the new microRNAs won't be annotated during the analysis.

Secondly, traditional tools for annotation and quantification genes from raw reads of total RNA-seq are designed for long RNA sequences annotation, which struggle to correctly identify microRNAs, due to size limitations or inadequate

alignment parameters for short sequences (Riaz et al., 2023). For example, TopHat and Cufflinks, which are commonly used for RNA-seq analysis for gene annotation, do not perform well when used to identify microRNA species due to the misalignment of short microRNA sequences (Fuller et al., 2024).

Thirdly, microRNA species are always in low expression when comparing to mRNA species, which lead to overshadowed of microRNA species by more abundant RNA species, especially in tools that do not have sufficient sensitivity for low-expression features (Yang et al., 2024).

2.3.1(b) Small RNA sequencing

Small RNAs, including microRNAs (microRNAs), play a vital role in the regulation of gene expression and have been shown to act as oncogenes or tumor suppressors in breast cancer (Chen et al., 2024). Specific sequencing technologies have been created for profiling the expression of small RNAs. MiR-seq is one of the critical tools that has enhanced the understanding of the roles of mature microRNA species in breast cancer (Riaz et al., 2023).

To conduct miR-seq, library of mature microRNA sequences is required. The library preparation including RNA isolation, size selection, ligation of adapter, reverse transcription and amplification by PCR. The following paragraphs introduce the process of each step.

RNA isolation is the process to isolate all RNA material from cell groups, tissues or other biological samples (Chomczynski & Sacchi, 1987). Operators during the isolation process must be careful enough to avoid the loss of microRNA species considering the average length of mature microRNA species is about 20 nucleotides

(nts) (Kim et al., 2012). Trizol method is the most widely used method for isolation of RNA species.

The size selection step ensures that there are only short fragments of transcription with the length between 18 -30 nts and all larger RNA species such as mRNAs are discarded (Technologies, 2010). To perform the size selection, different methods (Gel-based, magnetic bead-based and spin column-based) can be adopted. Gel-based size selection is the most precise methods, while the most time-consuming. (Liu et al., 2012) The other two methods run with high efficiency but precision is sacrificed.

As for the ligation of adapters, which adds adapters to 3' and 5' ends of the selected short RNA species with both single-strand RNAs by enzyme. The selected RNA species become recognized by ligation of adapters, makes it fulfil the requirement of sequencing library conduction (Hafner et al., 2008). However, the bias to specific sequences and secondary structures will be introduced to the library during the process of the ligation of the artificial adapters (Jayaprakash et al., 2011). Purification is compulsory after ligation of adapters to remove unnecessary reads.

Reversed transcription is the process to convert RNA species to cDNA reads to avoid the errors introduced by the process of RNA species degradation (Technologies, 2010). The process requires careful optimization of enzyme choice, primer design, and reaction conditions to ensure efficient and accurate cDNA synthesis. This cDNA serves as the template for PCR amplification and subsequent sequencing.

PCR amplification is the last step before sequencing. The aim of PCR amplification is to generate sufficient material for sequencing by amplicon of cDNA obtained from reverse transcription step.

Through miR-seq, researchers can profile microRNA expression and discover novel microRNAs that may have previously been overlooked. In breast cancer, microRNAs such as miR-21, miR-155, and miR-10b have been found to be significantly dysregulated and contribute to tumorigenesis, metastasis, and therapy resistance (Iorio & Croce, 2009). miR-21, for example, has been identified as an oncomiR that targets multiple tumor suppressor genes, promoting cancer cell survival and proliferation (Li et al., 2020).

The precise quantification of microRNAs through miR-seq has led to the identification of microRNAs that can serve as biomarkers for early breast cancer detection, as well as potential therapeutic targets for modulating gene expression in cancer cells. This technology has expanded the scope of breast cancer research, moving beyond protein-coding genes to explore the regulatory roles of small RNAs.

2.3.2 High-Throughput screening for microRNA-mRNA interactions

Understanding the interactions between microRNAs and their target mRNAs is crucial for elucidating the regulatory networks that govern breast cancer progression. High-throughput screening methods have been developed to systematically identify these interactions, allowing researchers to map microRNA-mRNA networks on a genome-wide scale.

Crosslinking immunoprecipitation followed by sequencing (CLIP-seq) and RNA immunoprecipitation (RIP)-seq can be both used to identify the interaction between microRNAs and target genes. CLIP-seq enables the identification of microRNA binding sites on mRNA transcripts by capturing the complex of RISC-microRNA-mRNA by specific antibodies (Chi et al., 2009). RIP-seq works similarly to CLIP-seq without crosslinking.