PREDICTIVE MOLECULAR BIOMARKERS FOR HUMAN HEALTH RISK

Guanglong Jiang

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Luddy School of Informatics, Computing, and Engineering
Indiana University

July 2024

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Yunlong Liu, Ph.D., Chair

Jingwen Yan, Ph.D.

May 15, 2024

Jun Wan, Ph.D.

Juexin Wang, Ph.D.

DEDICATION

To my father, whose unwavering support and boundless love have always been my guiding light. Your legacy continues to inspire me, and I carry your wisdom in every step. May you find eternal peace and watch over me from heaven. This achievement is as much yours as it is mine.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to all who supported me during my Ph.D. journey. Foremost, I am profoundly thankful to my advisor, Professor Yunlong Liu, for his continuous encouragement and invaluable mentorship, which were essential to my development and success in my studies. My sincere appreciation also extends to my committee members: Professor Jingwen Yan, Professor Jun Wan, and Professor Juexin Wang. Their insightful comments and suggestions have significantly shaped my research, providing clarity and direction.

I am indebted to Professor Bryan P. Schneider for his unreserved support, as well as to my colleagues Dr. Hongyu Gao, Dr. Fei Shen, Dr. Dr. Jill L. Reiter, Milan Radovich, Dr. Santosh Philips, and many others at the School of Medicine. Their assistance was crucial in advancing my work. I am also grateful for the camaraderie and wisdom shared by the faculty and staff members of the Luddy School of Informatics, Computing, and Engineering, which have been fundamental to my progress.

Lastly, my deepest thanks go to my family: My wife has been a cornerstone of support, managing our home with grace and resilience, and maintaining belief in me throughout this challenging journey. My children’s love and encouragement have been a constant source of joy and motivation. My parents have provided unwavering spiritual support, for which I am eternally grateful.

Guanglong Jiang

PREDICTIVE MOLECULAR BIOMARKERS FOR HUMAN HEALTH RISK

Molecular biomarkers play vital roles in disease risk assessment, personalized treatment selection and therapy response monitoring. This thesis explores the use of diverse molecular biomarkers for the assessment of human health risks, primarily in cancers. MiRNAs and their isoforms (isomiR) are promising biomarker candidates due to their comprehensive regulation of gene expression and involvement in physiology and pathological processes. The first study demonstrates that genetic variations in miRNA precursor regions influence the biogenesis of isomiRs in 95 SNP-isomiR pairs. Notably, we identified a SNP (rs6505162) impacting hsa-miR-423 isomiRs, potentially linked to breast cancer pathogenesis, suggesting their potential as biomarkers in disease assessment. The findings also highlight the mechanism of genetic regulation of isomiR generation and advance our understanding of miRNA mediated post-transcriptional regulation. Secondly, we explored the predictive capacity of aberrant intron-retention neoantigen burden (INB) in predicting the response to immune checkpoint inhibitors (ICI) in metastatic cancers. Both INB and tumor mutation burden (TMB) were strong predictors of ICI therapy duration (p = 0.019 and 0.038, respectively), with patients exhibiting elevated levels demonstrating exceptional treatment duration. Patients with high INB or TMB had improved overall survival (OS) (p = 1.1×10^{-4}). Importantly, INB and TMB were uncorrelated, indicating that they capture distinct aspects of tumor neoantigen. Together, the combined assessment of INB and TMB offers improved accuracy in predicting clinical response to ICI therapies. Finally, we extend the application of molecular biomarkers to

the assessment of minimal residual disease for risk stratification in triple negative breast cancer (TNBC) with residual disease after neoadjuvant chemotherapy. Detection of circulating tumor DNA (ctDNA) was a significant predictor of inferior distant disease-free survival (DDFS) ($p = 0.006$), disease-free survival (DFS) ($p = 0.009$) and OS ($p = 0.002$). The combination of circulating tumor cell (CTC) and ctDNA markers provided superior sensitivity and prognostic value. In conclusion, the studies provide compelling evidence for the utility of diverse molecular biomarkers – including miRNA isoforms, abnormal splicing-based neoantigen metrics and circulating tumor DNA in disease prediction and treatment efficacy assessment. By elucidating the roles of diverse biomarkers in predicting cancer pathogenesis and therapeutic response, we pave the way towards more personalized and effective approaches to managing human health risks.

Yunlong Liu, Ph.D., Chair

Jingwen Yan, Ph.D.

Jun Wan, Ph.D.

Juexin Wang, Ph.D.

TABLE OF CONTENTS

List of Tables x
List of Figures xi
List of Abbreviations xii
Chapter One: Introduction 1
Theme One: Genetic Regulation of IsomiR Biogenesis 3
1.1 Background 3
1.2 Objective 3
1.3 Significance 4
1.4 Innovation 5
Theme Two: Intron-retention Neoantigen Burden and Response to ICI therapy 6
2.1 Background 6
2.2 Objective 7
2.3 Significance 7
2.4 Innovation 7
Theme Three: ctDNA/CTC and Disease Recurrence in TNBC with Residual Disease 8
3.1 Background 8
3.2 Objective 8
3.3 Significance 8
3.4 Innovation 9
Chapter Two: Cis-Acting Genetic Variants and IsomiRs Biogenesis 11
2.1 Literature Review 11
2.2 Materials and Methods 13
2.2.1 Datasets 13
2.2.2 sRNA-seq Data Processing and IsomiR Identification 13
2.2.3 IsomiR Classification 14
2.2.4 Allele-specific Alignment 15
2.2.5 Genetic Association 16
2.3 Results 18
2.3.1 The 5'-end Base Nucleotide Substitution and IsomiR Stability 22
2.3.2 Canonical subtypes may not be the most abundant isomiRs 22
2.3.3 Allele-Specific Expression and Genetic Associations 23
2.4 Discussion 36
Chapter Three: Aberrant Intron-retention Derived Neoantigen Burden Predicts Response to Immune Checkpoint Inhibitor Therapy 40
3.1 Literature Review 40
3.2 Methods 44
3.2.1 Patient Population 44
3.2.2 RNA Sequencing and Data Processing 44
3.2.3 Prediction of Aberrant Intron-retention Neoantigen 45
3.2.4 Statistical Analysis 45
3.3 Results 47
3.3.1 Patient Data 47

3.3.2 Higher INB and TMB Levels in Patients with Long Duration of ICI Therapy 52
3.3.3 INB and TMB Were Independently Associated with ICI Therapy Duration 54
3.3.4 INB and TMB Predict ICI Therapy Duration in Cross Validation 60
3.3.5 INB and TMB Predict ICI Treatment-related Overall Survival 62
3.4 Discussion 64
Chapter Four: Circulating Tumor DNA in Prediction of Disease Recurrence in Triple-Negative Breast Cancer 68
4.1 Literature Review 68
4.2 Method 71
4.2.1 Clinical trial and sample acquisition 71
4.2.2 Sequencing and data processing 73
4.2.3 Sequencing and data processing 73
4.2.4 CTC isolation and enumeration 74
4.2.5 Statistical analysis 74
4.3 Results 75
4.3.1 Patient demographics and sample selection 75
4.3.2 Association of ctDNA With Survival Outcomes 79
4.3.3 Association of the Combination of CTCs and ctDNA With Clinical Outcomes 81
4.4 Discussion 85
References 88
Curriculum Vitae

LIST OF TABLES

Table 2.1 Correlation results between SNPs and isomiR compositions.....31
Table 2.2 Association of rs6505162 with hsa-miR-423 isomiRs 35
Table 3.1 Demographic and clinical variables.....50
Table 3.2 Summary of immune checkpoint inhibitors received and patient cancer types 51
Table 3.3 Logistic regression for duration of ICI therapy with aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) levels 57
Table 3.4 Summary of immune checkpoint inhibitors received and patient cancer types 59
Table 4.1 Clinical and pathological characteristics of patients analyzed for plasma ctDNA in the BRE12-158 trial 77

LIST OF FIGURES

Figure 2.1 IsomiR subtypes in human lymphoblastoid cell lines 20
Figure 2.2 Base quality of 5'-end nucleotide in different isomiR molecules 21
Figure 2.3 Allele-specific expression of hsa-miR-1304-3p for rs2155248 25
Figure 2.4 Scatter plot of the first two principal components.....26
Figure 2.5 IsomiR expression profile in human lymphoblastoid cell lines and TCGA breast cancer.....29
Figure 2.6 IsomiR expression in TCGA kidney renal clear cell carcinoma 30
Figure 3.1 Study Schema 43
Figure 3.2 CONSORT Diagram 49
Figure 3.3 Boxplot for aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) for short (<3 months) and long (>12 months) DoT patients 53
Figure 3.4 Scatter plot of aberrant intron-retention neoantigen burden (INB) and log-scaled tumor mutation burden (TMB) 55
Figure 3.5 Histogram for aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) 56
Figure 3.6 Receiver operating characteristic (ROC) and precision recall (PR) curves 61
Figure 3.7 Kaplan-Meier curves for overall survival grouped by aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) 63
Figure 4.1 BRE12-158 Trial Schema.....72
Figure 4.2 CONSORT diagram 76
Figure 4.3 Association of the presence of ctDNA with survival outcomes.....80
Figure 4.4 Association of the presence of CTCs with survival outcomes 82
Figure 4.5 Association of presence of ctDNA and CTCs with Survival Outcomes.....84

LIST OF ABBREVIATIONS

3’-UTR	3-prime untranslated region
AIC	Akaike information criterion
CEU	Northern Europeans from Utah
cfDNA	cell-free DNA
CLIA	clinical Laboratory Improvement Amendments
CONSORT	Consolidated Standards for Reporting Trials
CTC	circulating tumor cell
ctDNA	circulating tumor DNA, deoxyribonucleic acid
CTLA-4	cytotoxic T-lymphocyte associated protein 4
DDFS	distant disease-free survival
DFS	disease-free survival
DNA	deoxyribonucleic acid
DoT	duration on therapy
EBI	European Bioinformatics Institute
ER	estrogen-receptor
FDA	Food and Drug Administration
FDR	false discovery rate
FIN	Finnish in Finland
GDC	Genomic Data Commons
GRB	British in England and Scotland
HER2	human epidermal growth factor receptor 2
HLA	the human leukocyte antigen

HR	hazard ratio
ICI	Immune checkpoint inhibitor
IHC	Immunohistochemical
INB	aberrant intron retention-derived neoantigen burden
Iras	immune-related adverse events
IRB	Institutional Review Board
IUHPG	Indiana University Health Precision Genomics
MAF	minor allele frequency
MHC	the major histocompatibility complex
miRNA	micro-RNA, Ribonucleic acid
MRD	minimal residual disease
MSAF	maximum somatic allele frequency
MTB	molecular tumor board
NAC	neoadjuvant chemotherapy
NGS	Next generation sequencing
NSCLC	non-small cell lung cancer
OS	overall survival
PCA	principal component analysis
pCR	pathologic complete response
PD-1	programmed cell death protein 1
PD-L1	programmed death-ligand 1
PFS	progression-free survival
PR	progesterone-receptor

RCB	Residual Cancer Burden
QTL	quantitative trait loci
RNA	Ribonucleic Acid
SNP	single nucleotide polymorphism
sncRNA	small non-coding RNA
sRNA-seq	small RNA sequencing
TILs	tumor-infiltrating lymphocytes
TMB	tumor mutational burden
TMB-H	tumor mutational burden-high
TNBC	triple-negative breast cancer
TCGA	The Cancer Genome Atlas
TSI	Tuscans from Italy
VAF	variant allele frequency
YRI	Yoruba in Ibadan, Nigeria

Chapter One: Introduction

Biomarkers are invaluable tools used in clinical practice and medical research which allow for the accurate and reproducible measurement of substances or characteristics in living organisms. They serve as indicators of the status, progression, or risk of developing certain medical or physiological conditions or disorders. Existing biomarkers contribute significantly to clinical practice and medical research for the purpose of early medical diagnosis, disease severity evaluation, therapeutic selection and treatment effectiveness monitoring, and acceleration of drug development.

The ongoing development and refinement of biomarkers represents a dynamic area of investigation, with the hunting for more precise and reliable ones remaining paramount. This pursuit is particularly fueled by revolutionary innovations in technologies like next-generation sequencing (NGS) technology, which offer high-throughput quantification capabilities at decreasing costs. Progress in NGS opens new avenues for the integration of omics data (genomics, transcriptomics, epigenomics, and proteomics, etc.) into biomedical research, providing unprecedented opportunities and insights for biomarker discovery, and driving the demand for more accurate and reliable biomarkers that can guide clinical decision-making. Omics data integration has the potential to revolutionize personalized medicine, by the identification of patient-specific biomarkers that aides in development of more targeted and effective treatment plans, ultimately improving patient outcomes. As the research in this field continues to advance, the development of novel biomarkers holds great promise for transforming the future of healthcare.

MicroRNAs (miRNAs) are prime candidates for biomarker development due to their critical role in regulating gene expression. Variations in their mature sequence, known

as isomiRs, exhibit altered target specificity compared to canonical miRNAs, leading to diverse regulatory effects on target gene expression. Elucidating factors that govern isomiR biogenesis, such as genetic variants, can provide valuable insights into miRNA-mediated gene regulation and its implications for disease pathogenesis.

Additionally, understanding immune system's response to cancer therapy, particularly immune checkpoint inhibitors (ICIs), is crucial in clinics. While ICIs have significantly improved cancer treatment outcomes, patient responses vary greatly. Traditional biomarkers like tumor mutational burden (TMB) and programmed death-ligand 1 (PD-L1) expression have limitations in predicting responses to ICIs.

Lastly, triple-negative breast cancer (TNBC) represents a particularly aggressive form of breast cancer, and many patients have residual disease after initial therapy, which places them at high risk for recurrence. Accurately identifying individuals most likely to experience recurrence is essential for personalizing treatment strategies and optimizing patient outcomes, which allows minimizing unnecessary toxicities for patients with low risk of recurrence. However, existing tools for risk stratification have limitations. The presence of circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) may be employed to detect minimal residual disease and assess the likelihood of disease recurrence.

To address these challenges, this research focuses on three promising areas with the potential to identify biomarkers that predict human health risks. This research aspires to enrich the growing body of knowledge in the field of biomarker discovery. Ultimately, this endeavor seeks to empower clinicians with more accurate tools for personalized medicine and enhanced healthcare outcomes.

Theme One: Genetic Regulation of IsomiR Biogenesis

1.1 Background: Dysregulation of miRNAs is a hallmark of various human diseases including cancer, Alzheimer's disease, diabetes, and immune disorders. MiRNAs exert their influence by binding to the 3'-untranslated regions (3'-UTR) of target mRNA molecules and play a pivotal role in the regulation of cellular processes such as cell communication, proliferation, differentiation, and apoptosis. As a single miRNA molecule is capable of regulating hundreds to thousands of genes, miRNAs are promising biomarkers for disease diagnosis or prognosis, and provide options for medical intervention against diverse pathogenic conditions. Throughout miRNA maturation, various enzymes including the ribonucleases Drosha and Dicer, introduce variations in the mature miRNA sequence, which are known as isomiRs. These miRNA isoforms enriched the sequence variations within miRNAs, expanding the spectrum of post-transcriptional regulation of target mRNAs, yet the mechanism underlying their biogenesis remains largely unknown.

1.2 Objective: I hypothesize that a single nucleotide polymorphism (SNP) in the precursor miRNA sequence may affect the cleavage behaviors of enzymes such as Drosha and Dicer, leading to variations in the composition of isomiRs. By investigating the relationship between genetic variants and the prevalence of 5'-end isomiRs, this research aims to elucidate the isomiR variation through genetic regulation and uncover the impacts of cis-acting genetic variants on 5'-end isomiR variation. These genetic associations may shed light on the mechanisms underlying isomiR biogenesis and provide new clues for developing targeted therapy, by introducing genetic variants to modify in vivo isomiR composition. Specifically, my investigation aims to,

1) *Characterize isomiR expression profiles in human lymphoblastoid cell lines.*

By meticulously studying the expression profiles of isomiRs, I aim to decode the diverse and dynamic repertoire of miRNA isoforms within human lymphoblastoid cell lines. This exploration aims to provide a comprehensive understanding of isomiR diversity, pinpoint the prevalence of distinct isomiR subtypes, and shed light on allelic-specific expression of miRNAs.

2) *Elucidate how cis-acting genetic variants regulate isomiR biogenesis.*

This research aims to untangle the intricate network of cis-acting genetic variants that influence the biogenesis of isomiRs. By deciphering the regulatory mechanisms that govern this process, I seek to shed light on the interplay between genetic variations and the production of functionally distinct isomiRs in lymphoblastoid cells.

3) *Identify SNPs and isomiRs as potential biomarkers for complex diseases.*

Based on results from the previous steps, this part of the investigation aims to determine the potential of both genetic variants and isomiRs as biomarkers for intricate human diseases. By investigating the association of genetic variants and isomiR expression with human diseases using datasets from public domain, we aim to identify elements as potential biomarkers for complex human diseases.

4) *Developing of tools for unbiased isomiRs identification and classification.*

There are various public tools designed for isomiR identification from NGS data with similar algorithms. However, most of these tools are unable to accommodate genetic variants in DNA sequence, resulting in a biased estimate of the isomiR expression profiles. During my research, I will develop an integrated tool that can identify, quantify, and

classify allelic-specific expression of isomiRs from small-RNA sequencing data in an unbiased manner.

1.3 Significance: This research delves deep into the intricate mechanisms governing miRNA biogenesis and function. By elucidating the influence of genetic variants on isomiR generation, we can gain a better understanding of how these potent regulators impact gene expression and disease development. This knowledge has the potential to revolutionize our understanding of miRNA biology and pave the way for the development of novel therapeutic strategies that target specific isomiRs or their regulatory pathways.

The associations between SNPs and isomiRs, particularly within cancer contexts, could potentially lead to novel diagnostic and therapeutic strategies. With the understanding how genetic variations influence isomiR biogenesis, we can potentially develop novel therapeutic strategies that manipulate isomiR expression or function to combat disease. This could involve leveraging specific isomiRs as therapeutic agents or targeting genetic variants to influence the production of beneficial isomiRs.

This study explores the previously understudied realm of isomiRs and their potential contribution to disease pathology. By comprehensively analyzing isomiR diversity and its genetic determinants, we can significantly broaden our understanding of miRNA regulation and its role in human health and disease. The prevalence of non-canonical miRNAs and allele-specific expressions will contribute to the broader understanding of miRNA regulation and its genetic modulation.

1.4 Innovation: This study employs a customized reference method to comprehensively identify isomiRs across the genome. This method goes beyond traditional approaches and has the potential to reveal a wider range of isomiRs unbiasedly, offering a more complete

picture of isomiR diversity within cells. This study takes a comprehensive approach to investigate the genetic regulation of isomiR biogenesis, combining bioinformatics analysis, genetic associations, and expression profiling. The investigation of isomiR composition and its genetic determinants represents a novel angle in miRNA research.

Theme Two: Intron-retention Neoantigen Burden and Response to ICI Therapies

2.1 Background: Cancer immunotherapy has drastically changed the treatment of cancers and significantly improved patient survival and quality of life. However, the majority of patients do not benefit from immune checkpoint inhibitors, and some develop diverse levels of toxicities. It is in urgent need to develop biomarkers to predict which patient will benefit from the immunotherapies to optimize treatment strategies and reduce toxicities.

Tumor-specific neoantigens generated by genetic alteration in tumor cells and presented by major histocompatibility complex class I on the cell surface may be recognized by T cells and trigger an anti-cancer immune response in the host. High tumor mutation burden (TMB) is associated with ICI response in melanoma and solid tumors, and correlated to neoantigen load. Neoantigen presentation and recognition substantially affect the efficacy of immune checkpoint blockade treatment, and potentially serve as a predictive biomarker for identifying patients benefit most from immunotherapies.

Source of tumor neoantigens includes tumor somatic mutations, viral infection, gene rearrangement and alternative splicing. Intron-retention arose from splicing errors leads to retaining of introns in the mature mRNAs and may produce neoepitopes that never expressed in normal cells. Our group has reported high retention-induced neoantigen load correlated with unfavorable prognosis in multiple myeloma, and favorable prognosis in

pancreatic cancer. It is not clear how aberrant intron-retention neoantigen will affect immune response in cancer patients.

2.2 Objective: This research investigates the relationship between aberrant intron retention-derived neoantigen burden (INB) and the efficacy of ICI therapies across metastatic tumor types. The study seeks to assess whether INB, independently and in combination with TMB, can predict ICI therapeutic efficacy and overall survival in patients.

2.3 Significance: Predicting which patients will benefit from ICI therapy remains a challenge due to the limitations of currently used biomarkers. This research aims to address this challenge by identifying intron-retention neoantigen burden as a valuable novel biomarker for patient stratification. Combining INB with established biomarkers like TMB could provide clinicians with a more comprehensive estimation of a patient's overall neoantigen burden. This, in turn, could lead to more accurate predictions of ICI treatment response, ultimately enabling better patient selection for suitable life-saving therapy. This study expands the understanding of neoantigen presentation by exploring a novel resource beyond traditional somatic mutations. By investigating INB, we can gain valuable insights into how alternative splicing errors contribute to the neoantigen landscape. This knowledge can shed light on the intricate mechanisms by which tumors evade immune detection and potentially pave the way for the development of novel therapeutic strategies that target intron retention-derived neoantigens or enhance their presentation to the immune system.

2.4 Innovation: This study breaks new ground by investigating intron retention as a source of neoantigens for predicting ICI response in tumor panel. Traditionally, research has focused on TMB as the primary source of neoantigens. By exploring INB, we delve into a

previously understudied area with the potential to significantly improve our understanding of neoantigen generation and its impact on ICI efficacy.

Theme Three: ctDNA/CTC and Disease Recurrence in TNBC with Residual.

3.1 Background: Triple-negative breast cancer (TNBC) is an aggressive disease characterized by higher probability of relapse and poorer survival outcome following neoadjuvant therapy. Studies indicate that TNBC is a heterogeneous disease with distinct histological and molecular subtypes and varying therapeutic responses. Unfortunately, around two-thirds of patients with TNBC treated with neoadjuvant therapy have residual disease, placing them at high risk of relapse. Therefore, stratifying TNBC patients to guide personalized treatment decisions holds immense clinical importance.

ctDNA and CTC, detectable through liquid biopsies, can be used to detect the presence of minimal residual disease. These biomarkers could facilitate personalized therapy approaches, individualizing therapy with the best response and minimum toxicity for patients with TNBC.

3.2 Objective: My research seeks to decipher the significance of plasma ctDNA and CTC in accurately predicting recurrence and survival outcomes for triple-negative breast cancer patients with residual disease.

3.3 Significance: The identification of reliable markers for disease recurrence in TNBC patients' post-neoadjuvant chemotherapy has significant clinical implications. If ctDNA and CTC prove to be robust predictors of recurrence risk, it enables major improvements in patient care. By accurately identifying patients at low risk of recurrence, clinicians can personalize treatment plans, allowing low-risk patients to potentially avoid additional

rounds of chemotherapy. This translates to a significant improvement in quality of life by avoiding unnecessary harmful side effects, and reduces in the treatment burden on the healthcare system.

3.4 Innovation: This study builds upon previous research by conducting a comprehensively analyzing ctDNA (and potentially CTCs) in a relatively large, well-characterized cohort of TNBC patients. The large sample size enhances the generalizability of findings. Examining the predictive power of both ctDNA and CTCs offers a comprehensive evaluation of a patient's minimal residual disease and the risk of recurrence. The combined approach has the potential to refine current risk stratification models and lead to the development of new, more precise precision medicine tools.

Chapter 2 will delve into the intricate processes governing the biogenesis of isomiRs, a class of microRNAs with slight variations from their canonical sequences. We will explore the genetic underpinnings that influence the production of isomiRs, focusing on how cis-acting genetic variants may alter miRNA processing and lead to the formation of these unique molecules. Chapter 3 will examine the role of aberrant intron-retention neoantigen burden, investigating how the neoantigen production from both retention of introns and tumor somatic mutations affect the immune response to cancer ICI therapies. Chapter 4 will discuss the use of ctDNA and CTC as biomarkers for predicting disease recurrence in triple-negative breast cancer after neoadjuvant chemotherapy with residual disease. Each chapter will begin with a detailed literature review, providing a solid theoretical foundation and context for the studies. This will be followed by a description of the methods and materials used in that research, ensuring a comprehensive understanding of the experimental designs and analytical techniques. Results from the

studies will be illustrated clearly, providing both quantitative and qualitative data. Discussions in each chapter will interpret the results, drawing connections to existing research and potential clinical applications.

By addressing these intricate questions, my research aspires to contribute to the growing body of knowledge in the field of biomarker discovery. Ultimately, this pursuit seeks to empower clinicians with more precise tools for personalized medicine, thereby improving health outcomes for individuals and communities.

Chapter Two: Cis-Acting Genetic Variants and IsomiRs Biogenesis

2.1 Literature Review

MicroRNAs (miRNAs) are a class of endogenous small noncoding RNAs (sncRNAs) found in most eukaryotes and have been linked to almost every aspect of physiological processes. MiRNAs regulate approximately 60% of human protein-coding genes (Friedman et al., 2009), and their dysregulation is a hallmark of various human diseases, including cancer, Alzheimer's disease, diabetes, and immune disorders (Ardekani & Naeini, 2010; Li & Kowdley, 2012). MiRNAs bind to the 3'-untranslated regions (3'-UTR) of target mRNA molecules, leading to the repression of gene transcription in most cases and activation in some rare instances (Dweep et al., 2011). As a result, miRNAs play a pivotal role in the regulation of cellular processes, such as cell communication, proliferation, differentiation, and apoptosis (Hwang & Mendell, 2006). Since a single miRNA molecule can regulate hundreds to thousands of genes, miRNAs are promising new biomarkers for disease diagnosis and prognosis, offering ample options for medical intervention against diverse pathogenic conditions (Hanna et al., 2019). Since their first discovery in 1993 (Lee et al., 1993), 38,589 miRNA entries have been recorded in miRBase (v22.1), including 1881 pre-miRNAs and 2588 mature miRNAs for Homo sapiens (Kozomara et al., 2019).

MiRNAs exist in the genome within intergenic regions with their own promoters, as well as in the introns or exons of host protein-coding genes. The expression of isomiRs is regulated dynamically and displays tissue-specific patterns (Telonis et al., 2017). During miRNA maturation, various enzymes, including the ribonucleases Drosha and Dicer, can introduce variations in mature miRNA sequences, resulting in what are known as isomiRs.

Initially considered sequencing artifacts (Lee et al., 2010), isomiRs are now recognized to be produced by shifts in the cleavage sites at the 3'- or 5'-terminus, nucleotide substitutions along the entire molecule, or nucleotide additions at either end that deviate from the reference sequence (Tomasello et al., 2021). In addition, isomiRs derived from the same precursor can possess different seed sequences (two to seven bases at the 5'-end), which enables them to target different mRNAs or potentially interact with canonical miRNA in a cooperative or competitive manner (Desvignes et al., 2015; van der Kwast et al., 2020), thereby expanding their scope of post-transcriptional gene regulation.

Proposed explanations for isomiR heterogeneity are based on the structure of miRNA precursors (Starega-Roslan et al., 2011), precursor processing (Humphreys et al., 2012), and the AGO2 protein (Juvvuna et al., 2012). In addition, studies on genetic variants, especially those in the promoter regions of precursor transcripts, have identified miRNA expression quantitative trait loci (miR-QTLs) that regulate canonical miRNA expression (Borel et al., 2011; Budach et al., 2016; Huan et al., 2015; Lappalainen et al., 2013). However, the mechanism underlying the biogenesis of isomiRs remains largely unknown. In this study, we hypothesize that a single-nucleotide polymorphism (SNP) in the pre-miRNA sequence may affect the cleavage behaviors of enzymes like Drosha and Dicer, leading to variations in the composition of isomiRs. By investigating the relationship between genetic variants and the prevalence of 5'-end isomiRs, we aim to explain the isomiR variations through genetic regulation and uncover the impacts of cis-acting genetic variants on 5'-end isomiR variations. These genetic associations may shed light on the mechanisms overseeing isomiR biogenesis and provide new clues for developing targeted therapy by introducing genetic variants to modify in vivo isomiR composition.

2.2 Materials and Methods

2.2.1 Datasets

The small RNA sequencing (sRNA-seq) data and corresponding metadata for 452 unrelated human lymphoblastoid cell lines were obtained from the Geuvadis project, and downloaded from ArrayExpress in fastq format (https://www.ebi.ac.uk/arrayexpress/). The phase 3 genetic variants data (release 20130502, GRCh38) for 1000 Genomes samples were downloaded from the European Bioinformatics Institute (EBI) FTP website (ftp://ftp.1000genomes.ebi.ac.uk/). The reference sequences for precursor and mature miRNAs were downloaded from miRBase (version 22, GRCh38) and filtered to retain human sequences only. The mature miRNAs in miRBase ranged from 16 to 28 bases in length. The human whole genome reference sequence (GRCh38) was downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables/). The Cancer Genome Atlas (TCGA) miRNA-seq and clinical data were obtained from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/, version 27.0).

2.2.2 sRNA-seq Data Processing and IsomiR Identification

The small RNA-seq reads were trimmed for 3'-adaptor sequences (TGGAATTCTCGGGTGCCAAGGAACTC) using cutadapt (version 1.9.1). Reads shorter than 16 nucleotides (the minimum length of mature miRNA in miRBase) after trimming were excluded from further analysis. To identify isomiRs from the next-generation sequencing data, we employed the isomiRID software, which has demonstrated high sensitivity and specificity in a previous study (Amsel et al., 2017). The isomiRID pipeline (version 0.53) (de Oliveira et al., 2013) follows a multi-step approach to identify isomiRs. In the first step (round 0 in isomiRID), reads perfectly matched to the precursor

sequences were mapped. Unmapped reads were then filtered using the whole genome reference to exclude reads from other genomic regions. In the second step (round 1 in isomiRID), the remaining unmapped reads were then aligned to pre-miRNAs with one base mismatch to identify one-base substitution isomiRs. Reads with more than one-base substitution were not considered in our study. sRNA-seq reads not mapped were further subjected to up to 5 rounds of trimming where a single base on either the 5' or 3' end was removed at each round. The trimmed reads were mapped to the pre-miRNA reference to identify non-templated additions. Alignment was performed using Bowtie v1 with the parameters --norc -a -v [0|1] --best --strata. Mapped reads with lengths in the range of 16 to 28 nucleotides were retained for analysis. The output of isomiRID was a tab-delimited text file containing the small RNA-seq reads aligned to the reference pre-miRNA hairpin sequences (de Oliveira et al., 2013). The isomiRs identified by isomiRID were further filtered to only keep sequence reads detected in at least 10 subjects.

2.2.3 IsomiR Classification

IsomiR classification was conducted using our homemade Python program to compare the aligned sequences to mature miRNAs from miRbase, which were also aligned to their precursor sequences. The classification was based on variations at the 5'- and 3'-ends, and assigned to categories such as canonical, substitution, trimming, templated extension, non-templated addition, or new isomiR. The canonical subtype had the same sequence as the mature miRNA. New isomiRs were defined as reads with less than 10 bases overlap with canonical miRNAs. Trimming isomiRs had a shorter sequence, while extension isomiRs had longer sequences compared to canonical ones. Both subtypes

aligned perfectly to the reference pre-miRNA hairpin. In contrast, addition isomiRs differed from the reference precursor sequences at either or both termini.

In the case of nucleotide substitutions, the default mapping behavior of isomiRID was adopted and only one-base substitution in the sequence was take into account, considering the short length of isomiRs. The substitutions at the seed regions (positions 2-7 of a miRNA (Lewis et al., 2005)) are crucial for mRNA target recognition and were classified as 5'-seed-substitution (5sSub isomiRs). Substitutions at the 5' or 3' termini were classified as single-base substitution (5Sub or 3Sub isomiRs) or multiple-base substitution (5mSub or 3mSub isomiRs). The frequency of isomiR subtypes identified for each pre-miRNA, and isomiR subtypes across all pre-miRNAs were assessed and illustrated.

2.2.4 Allele-specific Alignment

The genomic coordinates for human miRNAs were obtained from miRBase (GRCh38). The genetic variants data from the 1000 Genomes Project were filtered to retain variants mapping to the precursor miRNA using VCFtools (version 0.1.13). Quality control was performed to keep subjects having sRNA-seq data, and common variants with minor allele frequency (MAF) greater than 1% among the study subjects.

By default, isomiRID uses the precursor sequences from miRBase as alignment references, and variations at the DNA level are not considered. This introduces biased alignments towards reference sequences and reads carrying alternative alleles were either mapped as single-base substitution or unmapped if they harbored more variants in the sequence. Possible solutions for the allele-specific alignment issues are, 1) allowing multiple mismatches in alignment; 2) masking the SNP positions with ambiguous letters;

and 3) personalized references alignment. Incorporating multiple mismatches reduces the precision of the short reads mapping. SNP-masking in the reference sequence reduced the reference allele bias but created bias toward one of the aligned alleles (Degner et al., 2009). In our study, a personalized reference alignment method was used where sequences incorporating SNP alleles were generated as reference sequences. sRNA-seq reads were aligned to both miRBase precursor sequences and sequences with alternative alleles. Reads mapped to these customized references were merged with reads mapped to canonical references, in which reads mapped in an earlier round of isomiRID were kept. For example, if a canonical isomiR with an alternative allele aligned to the miRBase reference in round one as a 1-base mismatch, but it mapped to the customized reference as perfect match at round zero, then only the perfectly matched alignment from round zero was kept. In this study, only SNPs used in genetic association studies were considered in the alignment.

2.2.5 Genetic Association

Genetic associations were conducted between the cis-acting SNPs and the ratio of 5'-end isomiR variants (substitution, trimming, extension, and non-template adding) against all isomiRs identified for that specific mature miRNA. The analysis focused on isomiRs with a 5'-terminus within ± 8 bases relative to canonical miRNAs. Reads having same length at the 5'-terminus as mature miRNA without substitution were considered as 5'-end canonical isomiRs. 5'-end substitution included reads with single or multi-base substitutions in the seed region. 5'-end trimming, extension, and addition were limited to modifications of a maximum 8 bases at the 5'-end and were further classified according to the number of base modifications. The observed counts were transformed by adding 0.5 to

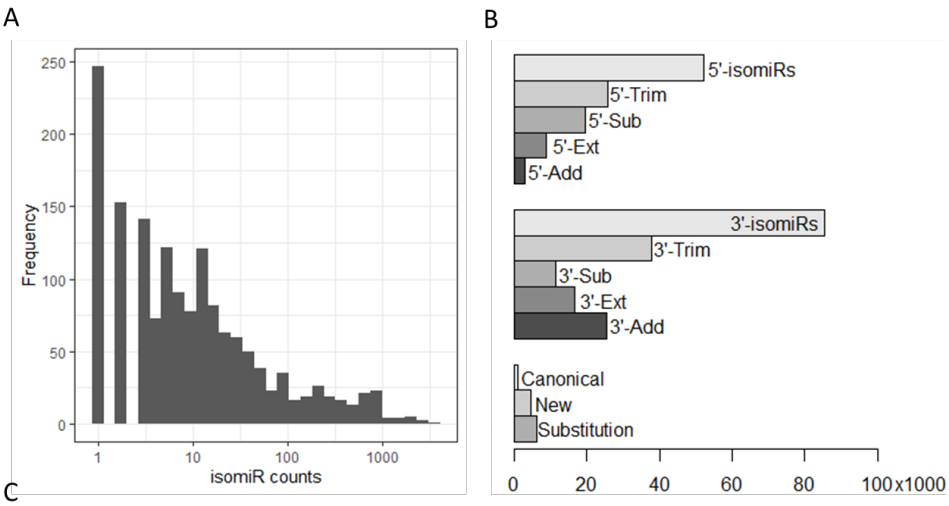
calculate the ratio. A non-parametric Kendal rank correlation (Kendall, 1938) was fitted using an additive genetic model where samples with none, one, or two rare alleles were coded as 0, 1, and 2. We conducted a principal component analysis (PCA) to detect and control for population stratification. The scatterplot of the first two principal components was used to illustrate the population stratification of the study subjects. The subjects included in the correlation analysis were from two populations, European Ancestry (Northern Europeans from Utah or CEU, Finnish in Finland or FIN, British in England and Scotland or GRB, Tuscans from Italy or TSI) and African Ancestry (Yoruba in Ibadan, Nigeria or YRI). To control for the population stratification, we conducted stratified Kendal rank correlation analyses with subjects from either the European or African population and required that the correlation be significant in overall correlation analysis and in correlation within either the European or African population. The Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was used to correct for multiple comparisons for 839 correlations across different subtypes of isomiRs, and the corrected p-value less than 0.05 was considered statistically significant.

To study the association between isomiR expression and tumor susceptibility, differential analysis of isomiRs between TCGA tumor and normal samples was conducted with isomiR expression data from the GDC Data Portal. The normalized reads (reads per million miRNAs mapped) for isomiRs with the same 5'-end variations were added together, and a student's t-test was used to assess the differences between tumor and normal samples. The statistical analyses and data visualization were conducted in statistical environment R (v3.6.1). *P*-value less than 0.05 was considered statistically significant.

2.3 Results

Our study aimed to investigate the genetic regulation of isomiRs, a type of variation in mature miRNA sequences, by characterizing their expression profiles in human lymphoblastoid cell lines. Published small RNA sequencing (sRNA-seq) data was gathered from 452 individuals and 652,778 isomiRs were mapped to 1,917 human precursor miRNA sequences. Quality filtering was applied to keep reads detected in at least 10 samples, resulting in 109,289 isomiRs mapping to 1,546 pre-mRNAs. The most abundant isomiR subtypes were found in hsa-mir-155 (3,501 isomiRs, ***Figure 2.1A***), which is consistent with the high levels of miR-155 expression in various human tissues and cell types and its multifunctional physiological roles (Faraoni et al., 2009). IsomiRs were classified according to variations at the 3'- or 5'-ends compared to a nearby mature miRNA (e.g., 5p or 3p miRNA). Consistent with previous reports (Neilsen et al., 2012; Woldemariam et al., 2019), we found that the variability in isomiR sequences occurred more frequently at the 3'-end, with 78% of isomiRs carrying variations at the 3'-end, 48% at the 5'-end, and more than one-third of isomiRs had variations at both ends (***Figure 2.1B***). Because sequencing library adaptor trimming during data preprocessing may affect variant calling at the 3'-end, we focused our study on 5'-end alterations, with base trimming being the most prevalent isomiR subtype (***Figure 2.1B***). Examples of 5'-end isomiR subtypes for hsa-miR-155 are shown in ***Figure 2.1C***. Non-templated nucleotide additions were also observed in the small-RNA-seq data. To discern whether these non-templated nucleotide additions might have arisen from sequencing errors, we compared the base quality of the 5' non-templated addition of hsa-miR-155-5p in ***Figure 2.1C*** against that of the templated extension across

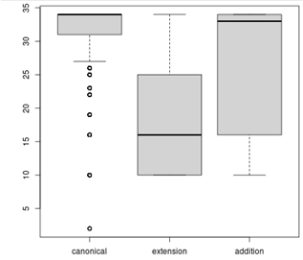
all study samples. We found that the base quality was similar for both the 5' non-templated addition and the templated extension (***Figure 2.2***).



-hsa-mir-155----------	5Var	3Var	isoType
#CUGUUAAUGCUAAUCGUGAUAGGGGUUUUUGCCUCCAACUGACUCCUACAUAUUAGCAUUAACAG			hsa-mir-155
#...UUAAUGCUAAUCGUGAUAGGGGUU.....			hsa-miR-155-5p
....UUAAUGCUAAUCGUGAUAGGGGUU.....	0	0	Canonical
......AAUGCUAAUCGUGAUAGGGGUU.....	-2	0	5Trim
....aTAATGCTAATCGTGATAGGGGTT.....	0	0	5Sub
....UUAAUaCUAAUCGUGAUAGGGGUU.....	0	0	5sSub
....gaAAUGCUAAUCGUGAUAGGGGUU.....	0	0	5mSub
...GUUAAUGCUAAUCGUGAUAGGGGUU.....	1	0	5Ext
.ugGUUAAUGCUAAUCGUGAUAGGGGUU.....	3	0	5Add
....UUAAUGCUAgUCGUGAUAGGGGUU.....	0	0	Sub
.....CUAAUCGUGAUAGGGGUUUUU.....	-6	3	5Trim||3Ext
...cUUAAUGCUAAUCGUGAUAGGGGUUUUU.....	1	3	5Add||3Ext
.....UUUGCCUCCAACUGACUCCU.....	NA	NA	New

Figure 2.1. IsomiR subtypes in human lymphoblastoid cell lines.

(A) Frequency of isomiRs identified for each precursor miRNA after quality filtering.
(B) Frequency of isomiR subtypes with 5'-end variations (total, trimming, substitution, extension, and addition), 3'-end variations, and other isomiR subtypes. The 5'-substitution (5'-Sub) category includes substitutions at the first base (5Sub), single-base substitutions at the seed region (5sSub), and 5' end multi-base substitutions (5mSub). The 3'-subcategory includes single-base (3Sub) or multi-base (3mSub) substitutions at the 3'-end. The Substitution bar (bottom) includes internal single-base substitutions.
(C) Examples of 5'-end isomiR subtypes for hsa-miR-155, including Canonical, 5'-trimming (5Trim), 5Sub, 5'sSub, 5mSub, 5'-extension (5Ext), 5'-addition (5Add), substitution in the middle (Sub), combinations of variants, and new isomiR. Variant nucleotides are shown in red. Lowercase letters indicate substitutions or non-templated additions. The seed sequence is underlined.



isomiR in fig 1C	isomiR sequence	BQ Mean	BQ SD
Canonical	UUAAUGCUAAUCGUGAUAGGGGUU	32.98	2.0
5' templated extension	GUUAAUGCUAAUCGUGAUAGGGGUU	18.21	7.6
5' non-templated addition	aUUAAUGCUAAUCGUGAUAGGGGUU	26.39	10.2

Figure 2.2. Base quality for 5'-end nucleotide in different isomiR molecules.

The boxplot and statistics (mean and standard deviation, SD) of the base quality (BQ) for the first nucleotide at the 5'-end of the hsa-miR-155-5p canonical, 5'- templated extension, and 5'- non-templated addition isomiRs across all study samples.

2.3.1. The 5’-end Base Nucleotide Substitution and IsomiR Stability

MiRNAs possess distinct half-lives in human cells, and miRNA stability is important for the dynamic regulation of cellular miRNA activity. Some of the factors that influence miRNA homeostasis include sequence modification, AGO protein complex formation, and mRNA target interaction (Kai & Pasquinelli, 2010; J. Wang et al., 2019). Furthermore, the nucleotide at the 5’-end was reported to influence mature miRNA stability, whereby miRNAs with an uracil (U) at the 5’-end, compared to guanine (G) or adenine (A), generally had significantly longer half-lives (Zhou et al., 2018). Therefore, we compared the frequency of the 5’-end bases between canonical and 5’-substitution isomiRs. Among 1,150 canonical miRNAs used for isomiR classifications, 42% contained a U at the 5’-end. In contrast, only 12.45% of substitution isomiRs had a U on the 5’-end. The most common end base in 5’-substitution isomiRs was G (39.21%), followed by A (24.90%) and cytosine (C) (23.44%). These results suggest that, in comparison to canonical miRNAs, 5’-end base substitution isomiRs are predicted to have shorter half-lives.

2.3.2. Canonical subtypes may not be the most abundant isomiRs

A previous study reported that a one-base shift isomiR of miR-140-3p (i.e., 1-base 5’-trim plus a 3’-extension) was expressed at higher levels compared to its canonical counterpart in breast cancer progression (Bhardwaj et al., 2018). We asked whether the human lymphoblastoid cell lines used in this study exhibited the same characteristic. To address this question, we compared read counts of the shifted isomiRs to the standard miR-140-3p miRNA, and discovered that the shifted isomiRs were more prevalent in 99% of the cells (448 out of 452). In addition, it is worth noting that canonical miRNAs annotated by miRBase were also not the predominant isomiRs for some of the miRNAs under

investigation. We compared the total reads across 452 subjects for a canonical miRNA and its isoforms in the 1000 Genomes sRNA-seq dataset for miRNAs with at least 5 subjects having 20 or more sequence reads. Non-canonical isomiRs were more abundant in 52% of the miRNAs. However, after considering the impact of 3'-adaptor trimming by grouping isomiRs based on sequence variations at the 5'-end, we observed that the frequency of miRNAs abundant in non-canonical isomiR categories was 11% of the total miRNAs. These findings imply that at least some non-canonical isomiRs might have important roles in miRNA regulation.

2.3.3. Allele-Specific Expression and Genetic Associations

Next, we asked whether cis-acting genetic variants impacted the frequency of isomiR subtypes. To address this question, we queried genotype data from the 1000 Genomes Project for SNPs located in DNA regions encoding precursor miRNA transcripts. We identified 4,478 variants, from which we selected 481 bi-allelic SNPs with a MAF greater than 1% in 435 individuals with both genotype and sRNA-seq data. We then used customized sequences with reference and alternative alleles for those SNPs to identify isomiRs transcribed from either allele. Our personalized references method also revealed allele-specific expression of miRNAs. As an example, the DNA sequence of hsa-miR-1304-3p has a SNP (rs2155248, T/G) at the 13th base (***Figure 2.3A***). Despite the T allele being the major allele among 435 individuals, we observed a higher number of miRNA reads with a C at the 13th nucleotide that were transcribed on the reverse strand of the G allele. This intriguing finding can be explained by the fact that cells heterozygous for rs2155248 (T/G) predominantly expressed the C allele isomiRs, while cells homozygous for T/T transcribed hsa-miR-1304-3p at very low levels (***Figure 2.3 B,C***). In addition,

allele-specific expression of miRNAs was commonly observed in the human lymphoblastoid cell lines where miRNAs containing alleles different from those annotated in miRBase were transcribed at lower levels compared to canonical miRNAs.

To investigate the genetic association of isomiR variants, we initially utilized the data from all study populations and conducted 839 associations between SNPs and different 5'-isomiRs. To account for population stratification, we conducted separate association studies with subjects of European or African ancestry populations (***Figure 2.4***). We set a criterion that the findings from all populations should be significant in either the European or African population study.

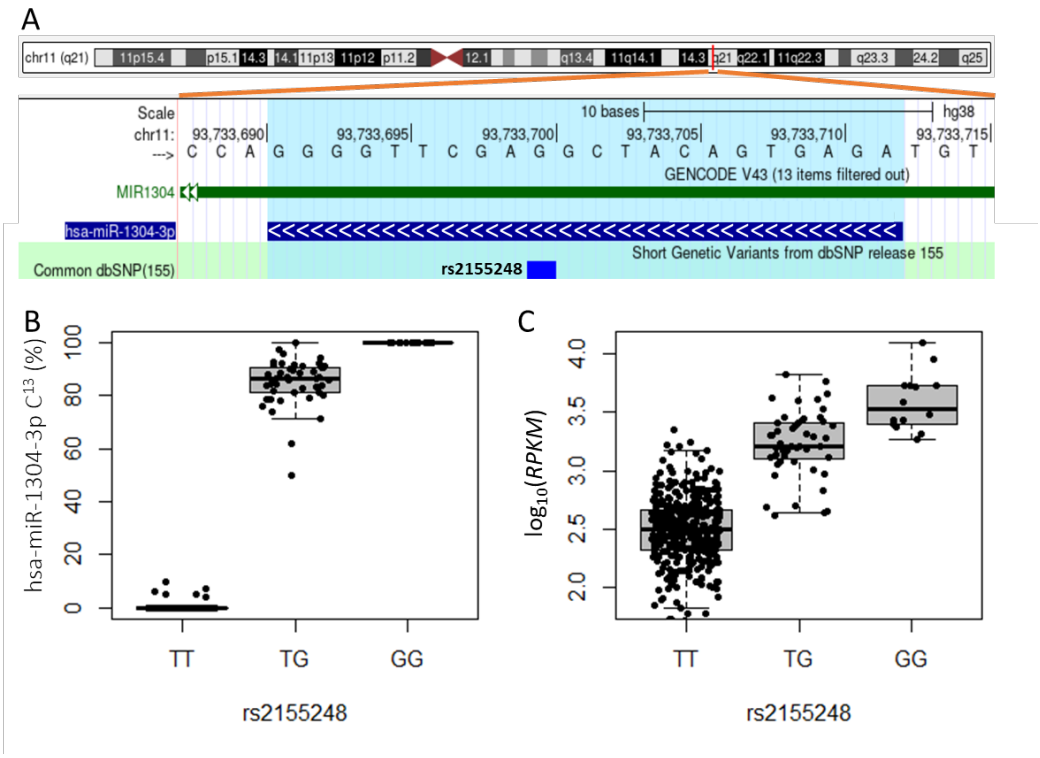


Figure 2.3. Allele-specific expression of hsa-miR-1304-3p for rs2155248.

(A) Schematic diagram showing the locations of the SNP rs2155248 and hsa-miR-1304-3p on chromosome 11q21. miRNA1304 is on the reverse strand.
(B) Percentage of hsa-miR-1304-3p miRNAs with cytosine at the 13th nucleotide location (C13) of rs2155248 in heterozygous and homozygous lymphoblastoid cell lines. The heterozygous (T/G) cells predominantly expressed miRNAs from the G allele. The cytosine reads in T/T genotype samples could arise from either single-nucleotide substitutions or sequencing errors.
(C) Expression levels of miR1304-3p among rs2155248 genotyped lymphoblastoid cell lines. Homozygous T/T cells transcribed low levels of the miRNA.

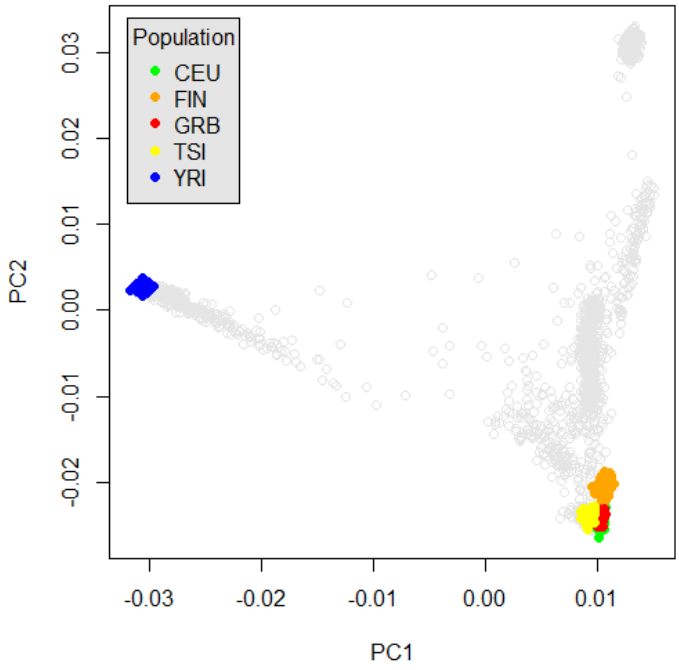


Figure 2.4. Scatter plot of the first two principal components.

1000 Genomes Project subjects included in this study are highlighted in colors,

- CEU: Utah residents with Northern and Western European Ancestry, European Ancestry, green
- FIN: Finnish in Finland, European Ancestry, orange
- GRB: British in England and Scotland, European Ancestry, red
- TSI: Toscani in Italy, European Ancestry, yellow
- YRI: Yoruba in Ibadan, Nigeria, African Ancestry, blue

Ultimately, we identified a total of 7, 51, 28, and 9 SNP–isomiR pairs that exhibited significance for 5'-substitution, -trimming, -extension, or -addition, respectively, using a threshold of false discovery rate (FDR) < 0.05 (***Table 2.1***). One notable finding was that rs6505162 (A/C) was associated with a two-base 5'-extension isomiR of hsa-miR-423-3p (FDR = 3.0×10^{-21}) and a two-base 5'-trimming isomiR of hsa-miR-423-5p (FDR = 2.4×10^{-17}), where the C allele was linked to a decrease in the expression of both isomiRs (**Table 2.2** and ***Figure 2.5A,B***). Interestingly, rs6505162 has been reported to have a high frequency of somatic mutation in breast cancer cell lines and tumor tissues (Zhao et al., 2015), and miR-423 activity was increased in breast cancer cells (Kontorovich et al., 2010). Therefore, to investigate the association between the hsa-miR-423 isomiRs and breast cancer pathology, we compared the expression levels of these isomiRs between tumor and normal samples in TCGA breast cancer (BRCA) dataset. We found that the expression of the two-base 5'-extension isomiR of hsa-miR-423-3p was significantly higher in tumor compared with normal tissue (p=0.00023, ***Figure 2.5C***) across all tumor subtypes and ethnic backgrounds, and in a subset of hormone-receptor-positive patients (estrogen-receptor-positive, ER+, or progesterone-receptor-positive, PR+) in the white population (p=0.0047). This finding is consistent with a prior study that reported that miR-423-3p, as compared with miR-423-5p, promoted cell proliferation and tumorigenesis in breast cancer (Zhao et al., 2015). However, we also observed higher expression of the 5'-trimming isomiR of has-miR-423-5p in tumors (p=0.00095, ***Figure 2.5D***), although the difference was not significant in the subset of hormone-receptor-positive white patients (p=0.056). To investigate whether the hsa-miR-423-3p two-base 5'-extension isomiR was associated with other cancers, we repeated our analysis using the TCGA kidney renal clear cell carcinoma

(TCGA-KIRC) dataset. We observed a similar finding that this isomiR was expressed at higher levels in tumors compared with normal tissues (p=0.00096, ***Figure 2.6A***), while a trend of high expression in tumors was also observed for hsa-miR-423-5p 5'-trimming isomiRs (p=0.064, ***Figure 2.6B***).

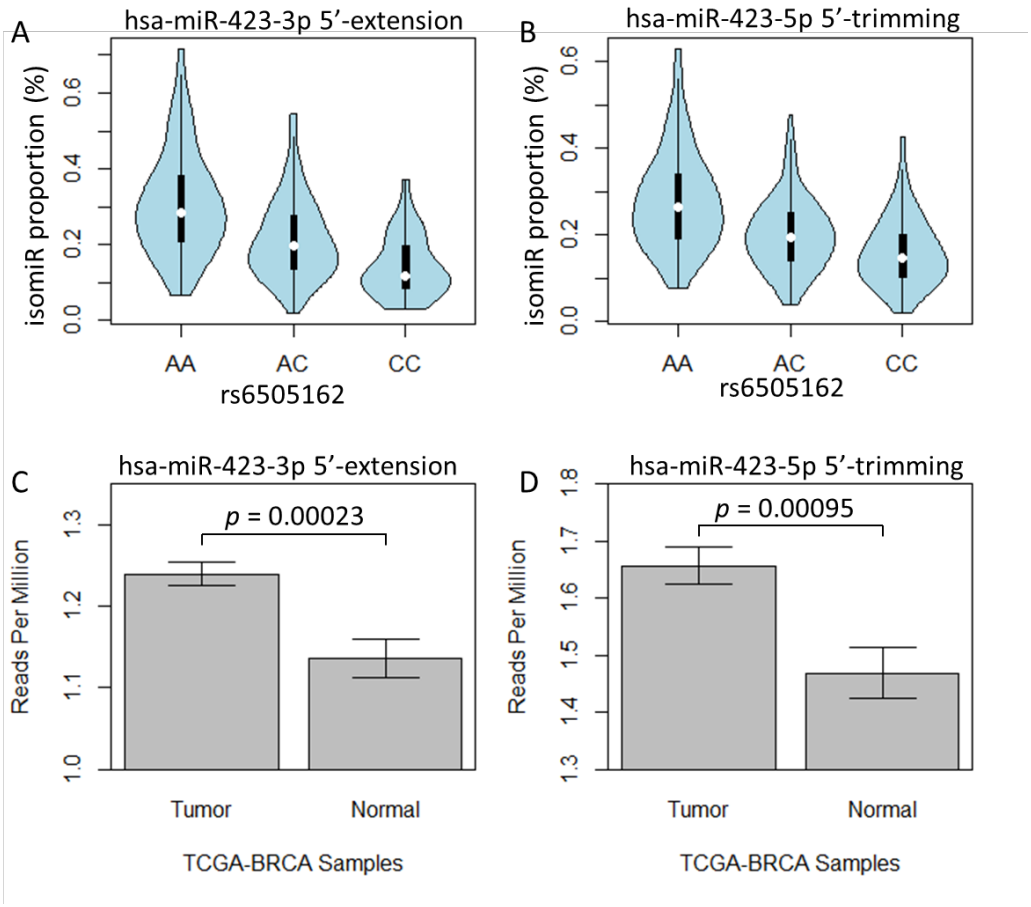


Figure 2.5. IsomiR expression profile in human lymphoblastoid cell lines and TCGA breast cancer.

Violin plot for (A) the proportion of hsa-miR-423-3p 5'-extension isomiRs and (B) the proportion of hsa-miR-423-5p 5'-trimming isomiRs in rs6505162 genotyped cells. Bar-plot for average expression level and standard error bar for (C) hsa-miR-423-3p 5'-extension isomiRs and (D) hsa-miR-423-5p 5'-trimming isomiRs in TCGA breast cancer tumor and normal samples.

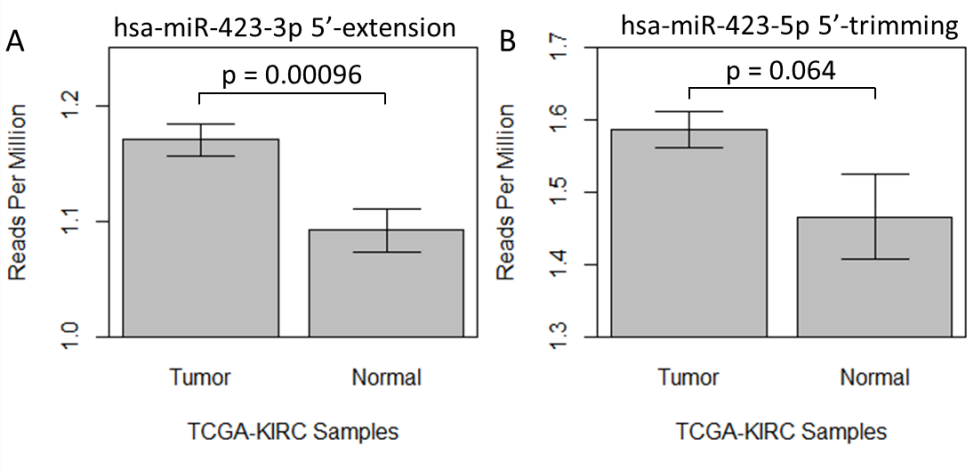


Figure 2.6. IsomiR expression in TCGA kidney renal clear cell carcinoma.

Bar-plot of average expression and standard error for **(A)** hsa-miR-423-3p 5’-extension isomiRs and **(B)** hsa-miR-423-5p 5’-trimming isomiRs in TCGA kidney renal clear cell carcinoma (TCGA-KIRC) and normal samples.

Table 2.1. Correlation results between SNPs and isomiR compositions.

Correlations with FDR < 0.05 in all subjects, and FDR < 0.05 in either European or African subjects.

SNP	isomiR				All(n=435)			European(n=348)			African(n=87)		
rsID	miRNA	Sub-type	#iso-miR	#base change	Tau	p-value	FDR	Tau	p-value	FDR	Tau	p-value	FDR
rs641071	hsa-miR-4482-3p	5-sub	1	0	-0.64	5.4E-60	4.5E-57	-0.59	7.0E-39	3.4E-36	-0.30	7.8E-04	0.012
rs641071	hsa-miR-4482-3p	5-trim	1	-1	-0.63	5.7E-59	2.4E-56	-0.59	1.0E-38	3.4E-36	-0.24	0.0054	0.066
rs641071	hsa-miR-4482-3p	5-trim	1	-2	-0.63	8.8E-59	2.5E-56	-0.59	1.8E-38	4.0E-36	-0.21	0.016	0.15
rs641071	hsa-miR-4482-3p	5-ext	1	1	-0.63	4.7E-58	9.8E-56	-0.57	1.6E-36	2.6E-34	-0.29	0.0012	0.017
rs2273626	hsa-miR-4707-3p	5-ext	1	1	0.50	1.0E-34	1.7E-32	0.51	2.4E-27	2.7E-25	0.20	0.035	0.26
rs2155248	hsa-miR-1304-3p	5-trim	3	-5	-0.47	3.1E-33	4.3E-31	-0.13	0.0025	0.024	-0.59	4.2E-12	2.9E-09
rs2155248	hsa-miR-1304-3p	5-trim	2	-4	-0.46	6.7E-33	7.4E-31	-0.13	0.0025	0.024	-0.58	8.7E-12	3.0E-09
rs34115976	hsa-miR-577	5-ext	15	1	-0.47	7.1E-33	7.4E-31	-0.49	9.5E-29	1.2E-26	-0.17	0.057	0.36
rs2168518	hsa-miR-4513	5-add	3	4	0.44	2.7E-30	2.5E-28	0.37	6.9E-17	3.3E-15	NA	NA	NA
rs2155248	hsa-miR-1304-3p	5-ext	24	2	0.44	5.4E-30	4.5E-28	0.15	6.2E-04	0.0077	0.57	1.5E-11	3.6E-09
rs2155248	hsa-miR-1304-3p	5-trim	6	-3	-0.43	4.1E-28	3.1E-26	-0.11	0.014	0.098	-0.49	9.7E-09	1.4E-06
rs487571	hsa-miR-5680	5-ext	9	2	0.40	3.1E-26	2.1E-24	0.39	3.6E-20	2.6E-18	0.41	1.2E-06	8.5E-05
rs10061133	hsa-miR-449b-5p	5-sub	2	0	-0.46	1.5E-23	1.0E-21	-0.48	3.1E-20	2.6E-18	-0.30	0.0051	0.062
rs6505162	hsa-miR-423-3p	5-ext	28	2	-0.37	4.9E-23	3.0E-21	-0.33	2.2E-15	9.6E-14	-0.20	0.025	0.20
rs28664200	hsa-miR-1255a	5-trim	1	-1	0.38	1.7E-21	9.5E-20	0.42	7.7E-21	7.2E-19	0.25	0.0050	0.062
rs6505162	hsa-miR-423-5p	5-trim	20	-2	-0.33	4.6E-19	2.4E-17	-0.30	6.5E-13	2.4E-11	-0.37	2.2E-05	9.0E-04
rs2155248	hsa-miR-1304-3p	5-ext	22	1	0.34	6.3E-19	3.0E-17	0.14	0.0018	0.019	0.50	3.0E-09	5.2E-07
rs7208391	hsa-miR-6868-3p	5-ext	2	1	0.38	6.4E-19	3.0E-17	0.34	2.6E-13	1.0E-11	0.45	6.1E-06	3.3E-04
rs7911488	hsa-miR-1307-3p	5-trim	131	-1	-0.33	3.2E-18	1.4E-16	-0.32	3.3E-14	1.4E-12	-0.10	0.25	0.78
rs2155248	hsa-miR-1304-3p	5-trim	54	-1	-0.33	2.7E-17	1.1E-15	-0.11	0.0095	0.077	-0.32	1.9E-04	0.0039
rs8054514	hsa-miR-3176	5-ext	2	1	0.31	8.1E-16	3.2E-14	0.26	4.0E-09	1.3E-07	0.46	6.6E-08	6.6E-06
rs75538180	hsa-miR-1303	5-ext	26	1	0.28	5.4E-14	2.0E-12	0.25	2.2E-09	7.2E-08	0.32	2.0E-04	0.0039
rs8054514	hsa-miR-3176	5-add	1	3	0.29	5.4E-14	2.0E-12	0.25	1.3E-08	3.4E-07	0.47	3.2E-08	3.7E-06

SNP	isomiR				All(n=435)			European(n=348)			African(n=87)		
rsID	miRNA	Sub-type	#iso-miR	#base change	Tau	p-value	FDR	Tau	p-value	FDR	Tau	p-value	FDR
rs75538180	hsa-miR-1303	5-ext	4	2	-0.28	6.6E-14	2.3E-12	-0.22	1.5E-07	3.2E-06	-0.22	0.011	0.11
rs77639117	hsa-miR-576-5p	5-trim	4	-1	0.29	9.0E-14	3.0E-12	0.26	4.4E-09	1.3E-07	0.41	3.1E-06	1.8E-04
rs8054514	hsa-miR-3176	5-trim	3	-1	0.28	4.8E-13	1.6E-11	0.25	1.8E-08	4.5E-07	0.44	2.4E-07	2.1E-05
rs174561	hsa-miR-1908-5p	5-trim	1	-1	-0.26	2.8E-11	8.8E-10	-0.29	3.2E-11	1.1E-09	-0.15	0.094	0.49
rs7227168	hsa-miR-4741	5-ext	1	3	0.26	1.1E-10	3.2E-09	0.26	1.8E-08	4.5E-07	0.32	4.3E-04	0.0071
rs487571	hsa-miR-5680	5-trim	2	-8	0.24	5.9E-10	1.7E-08	0.22	1.4E-07	2.9E-06	0.23	0.0069	0.075
rs487571	hsa-miR-5680	5-trim	2	-3	0.22	8.2E-09	2.3E-07	0.20	3.9E-06	7.0E-05	0.30	3.9E-04	0.0068
rs487571	hsa-miR-5680	5-ext	1	3	0.22	1.0E-08	2.8E-07	0.21	1.0E-06	1.9E-05	0.22	0.011	0.11
rs8054514	hsa-miR-3176	5-sub	6	0	0.22	1.1E-08	2.9E-07	0.17	6.7E-05	0.0011	0.41	1.7E-06	1.1E-04
rs2682818	hsa-miR-618	5-trim	1	-6	0.22	1.3E-08	3.3E-07	0.25	1.2E-08	3.4E-07	0.34	7.9E-05	0.0022
rs74085143	hsa-miR-4781-3p	5-trim	5	-2	0.23	1.3E-08	3.3E-07	0.14	0.0018	0.019	0.39	1.2E-05	5.7E-04
rs487571	hsa-miR-5680	5-ext	1	7	0.21	1.7E-08	4.0E-07	0.21	1.3E-06	2.4E-05	0.20	0.019	0.17
rs487571	hsa-miR-5680	5-ext	2	5	0.20	7.5E-08	1.7E-06	0.19	7.9E-06	1.3E-04	0.22	0.0089	0.096
rs2682818	hsa-miR-618	5-trim	1	-7	0.21	8.0E-08	1.8E-06	0.24	3.7E-08	9.0E-07	0.32	2.0E-04	0.0039
rs487571	hsa-miR-5680	5-trim	1	-4	0.20	8.0E-08	1.8E-06	0.19	6.6E-06	1.1E-04	0.21	0.012	0.12
rs2682818	hsa-miR-618	5-trim	2	-1	0.21	8.7E-08	1.9E-06	0.24	4.0E-08	9.2E-07	0.35	6.8E-05	0.0022
rs2682818	hsa-miR-618	5-trim	2	-3	0.21	9.7E-08	2.0E-06	0.24	3.9E-08	9.1E-07	0.32	2.9E-04	0.0054
rs4822739	hsa-miR-548j-5p	5-sub	2	0	-0.22	1.0E-07	2.1E-06	-0.11	0.018	0.12	-0.40	8.6E-06	4.3E-04
rs487571	hsa-miR-5680	5-trim	6	-1	0.20	2.0E-07	3.9E-06	0.17	8.4E-05	0.0013	0.32	1.5E-04	0.0036
rs7911488	hsa-miR-1307-3p	5-trim	2	-6	0.20	2.1E-07	4.1E-06	0.15	5.0E-04	0.0064	0.12	0.16	0.63
rs9589207	hsa-miR-92a-1-5p	5-trim	5	-2	-0.20	2.7E-07	5.1E-06	NA	NA	NA	-0.43	1.1E-06	8.5E-05
rs5997893	hsa-miR-3928-3p	5-trim	33	-2	-0.20	2.7E-07	5.1E-06	-0.13	0.0025	0.024	-0.11	0.21	0.70
rs2682818	hsa-miR-618	5-trim	2	-5	0.20	3.5E-07	6.3E-06	0.23	2.5E-07	5.0E-06	0.36	4.3E-05	0.0015
rs2682818	hsa-miR-618	5-trim	2	-4	0.19	8.1E-07	1.4E-05	0.23	1.4E-07	2.9E-06	0.30	5.2E-04	0.0085
rs151318590	hsa-miR-500b-3p	5-ext	2	2	-0.22	1.1E-06	2.0E-05	NA	NA	NA	-0.42	1.7E-05	7.3E-04
rs4822739	hsa-miR-548j-5p	5-add	2	1	-0.20	1.4E-06	2.3E-05	-0.09	0.048	0.26	-0.37	4.4E-05	0.0015

SNP	isomiR				All(n=435)			European(n=348)			African(n=87)		
rsID	miRNA	Sub-type	#iso-miR	#base change	Tau	p-value	FDR	Tau	p-value	FDR	Tau	p-value	FDR
rs9589207	hsa-miR-92a-1-5p	5-trim	3	-3	-0.19	1.7E-06	2.8E-05	NA	NA	NA	-0.33	1.6E-04	0.0036
rs151318590	hsa-miR-500b-3p	5-trim	1	-4	-0.21	1.8E-06	2.9E-05	NA	NA	NA	-0.40	3.4E-05	0.0013
rs2682818	hsa-miR-618	5-trim	3	-2	0.19	1.9E-06	3.0E-05	0.22	4.9E-07	9.5E-06	0.30	5.6E-04	0.0089
rs7911488	hsa-miR-1307-3p	5-trim	2	-7	0.18	2.4E-06	3.6E-05	0.14	8.2E-04	0.0096	0.05	0.58	0.88
rs74085143	hsa-miR-4781-3p	5-trim	3	-1	0.19	2.7E-06	4.0E-05	0.11	0.020	0.13	0.35	7.8E-05	0.0022
rs9589207	hsa-miR-92a-1-5p	5-trim	3	-1	-0.18	3.0E-06	4.4E-05	NA	NA	NA	-0.34	1.0E-04	0.0026
rs74085143	hsa-miR-4781-3p	5-trim	1	-4	0.19	4.3E-06	6.2E-05	0.12	0.0095	0.077	0.35	9.3E-05	0.0025
rs1439619	hsa-miR-3175	5-add	1	4	-0.20	4.5E-06	6.3E-05	-0.22	7.7E-06	1.3E-04	-0.13	0.18	0.65
rs4822739	hsa-miR-548j-5p	5-ext	1	8	-0.19	5.1E-06	7.2E-05	-0.08	0.082	0.34	-0.36	7.8E-05	0.0022
rs35613341	hsa-miR-5189-5p	5-trim	1	-1	0.20	6.5E-06	9.0E-05	0.21	4.0E-05	6.4E-04	0.20	0.040	0.27
rs4822739	hsa-miR-548j-5p	5-add	1	2	-0.18	1.2E-05	1.5E-04	-0.08	0.080	0.34	-0.35	1.3E-04	0.0032
rs10422347	hsa-miR-4745-5p	5-add	1	2	0.18	1.3E-05	1.6E-04	0.18	9.2E-05	0.0014	0.13	0.14	0.59
rs4534339	hsa-miR-1843	5-ext	29	1	-0.17	1.3E-05	1.6E-04	-0.09	0.034	0.20	-0.29	7.7E-04	0.012
rs9589207	hsa-miR-92a-1-5p	5-ext	1	1	-0.17	1.4E-05	1.8E-04	NA	NA	NA	-0.28	0.0013	0.018
rs4822739	hsa-miR-548j-3p	5-trim	1	-2	-0.20	1.6E-05	2.0E-04	-0.05	0.39	0.67	-0.38	1.9E-04	0.0039
rs7911488	hsa-miR-1307-3p	5-add	4	1	0.16	1.6E-05	2.0E-04	0.15	4.0E-04	0.0054	0.02	0.86	0.98
rs9589207	hsa-miR-92a-1-5p	5-trim	1	-4	-0.17	1.8E-05	2.2E-04	NA	NA	NA	-0.28	0.0018	0.024
rs151318590	hsa-miR-500b-3p	5-trim	1	-3	-0.19	2.0E-05	2.3E-04	NA	NA	NA	-0.34	4.0E-04	0.0068
rs2620381	hsa-miR-627-5p	5-trim	6	-1	-0.17	2.1E-05	2.5E-04	-0.06	0.17	0.46	-0.32	3.9E-04	0.0068
rs151318590	hsa-miR-500b-3p	5-ext	1	1	-0.19	2.4E-05	2.8E-04	NA	NA	NA	-0.35	3.2E-04	0.0059
rs9589207	hsa-miR-92a-1-5p	5-sub	21	0	-0.16	3.1E-05	3.5E-04	NA	NA	NA	-0.33	1.9E-04	0.0039
rs56292801	hsa-miR-5189-5p	5-trim	1	-1	0.19	3.9E-05	4.4E-04	0.18	3.0E-04	0.0043	0.19	0.060	0.37
rs9589207	hsa-miR-92a-1-5p	5-add	1	1	-0.16	4.9E-05	5.4E-04	NA	NA	NA	-0.28	0.0014	0.019
rs2427556	hsa-miR-941	5-add	2	1	0.15	1.1E-04	0.0011	0.15	4.0E-04	0.0054	0.33	7.0E-05	0.0022
rs10422347	hsa-miR-4745-5p	5-trim	1	-4	0.15	1.3E-04	0.0014	0.15	7.8E-04	0.0094	0.13	0.16	0.63
rs10422347	hsa-miR-4745-5p	5-ext	2	1	0.15	2.8E-04	0.0029	0.14	0.0015	0.016	0.11	0.21	0.71

SNP	isomiR				All(n=435)			European(n=348)			African(n=87)		
rsID	miRNA	Sub-type	#iso-miR	#base change	Tau	p-value	FDR	Tau	p-value	FDR	Tau	p-value	FDR
rs6413505	hsa-miR-6886-5p	5-trim	2	-2	0.16	3.4E-04	0.0034	0.07	0.19	0.49	0.38	1.5E-04	0.0036
rs2854001	hsa-miR-6891-5p	5-ext	18	1	-0.16	4.0E-04	0.0040	-0.16	0.0011	0.012	-0.04	0.70	0.91
rs74085143	hsa-miR-4781-3p	5-sub	1	0	0.14	4.6E-04	0.0045	0.11	0.020	0.13	0.26	0.0036	0.046
rs13186787	hsa-miR-1294	5-trim	2	-1	-0.14	4.9E-04	0.0047	-0.16	3.8E-04	0.0053	NA	NA	NA
rs75715827	hsa-miR-944	5-ext	8	1	-0.14	5.0E-04	0.0047	-0.14	0.0013	0.014	NA	NA	NA
rs4534339	hsa-miR-1843	5-ext	1	5	0.13	5.8E-04	0.0053	0.07	0.13	0.42	0.25	0.0034	0.044
rs13186787	hsa-miR-1294	5-ext	3	1	-0.14	6.1E-04	0.0055	-0.15	7.2E-04	0.0087	NA	NA	NA
rs12314280	hsa-miR-5700	5-trim	2	-2	0.15	8.7E-04	0.0077	0.18	4.5E-04	0.0059	0.04	0.68	0.91
rs78979347	hsa-miR-9903	5-sub	2	0	-0.15	0.0011	0.0092	-0.17	5.8E-04	0.0073	NA	NA	NA
rs4534339	hsa-miR-1843	5-trim	2	-3	0.13	0.0013	0.012	0.06	0.18	0.48	0.25	0.0034	0.044
rs10422347	hsa-miR-4745-5p	5-trim	6	-2	0.13	0.0016	0.014	0.13	0.0034	0.032	0.05	0.55	0.87
rs191321849	hsa-miR-500a-3p	5-ext	29	1	0.12	0.0019	0.015	0.13	0.0026	0.025	NA	NA	NA
rs191321849	hsa-miR-500a-3p	5-trim	4	-4	0.12	0.0025	0.020	0.12	0.0047	0.043	NA	NA	NA
rs174561	hsa-miR-1908-3p	5-ext	16	1	-0.12	0.0027	0.021	-0.16	2.2E-04	0.0032	-0.03	0.73	0.93
rs75258105	hsa-miR-5700	5-trim	2	-2	0.14	0.0029	0.022	0.16	0.0015	0.016	NA	NA	NA
rs515924	hsa-miR-548al	5-trim	1	-1	-0.12	0.0036	0.027	-0.14	0.0027	0.026	-0.07	0.48	0.87
rs191321849	hsa-miR-500a-3p	5-trim	8	-1	0.11	0.0036	0.027	0.12	0.0049	0.043	NA	NA	NA
rs6413505	hsa-miR-6886-5p	5-trim	3	-1	0.13	0.0052	0.036	0.06	0.26	0.56	0.32	0.0014	0.019
rs78979347	hsa-miR-9903	5-trim	2	-1	-0.12	0.0055	0.037	-0.15	0.0018	0.019	NA	NA	NA
rs13186787	hsa-miR-1294	5-ext	2	3	-0.11	0.0069	0.044	-0.13	0.0048	0.043	NA	NA	NA

Table 2.2. Association of rs6505162 with hsa-miR-423 isomiRs.

Negative Kendall rank correlation coefficients (Tau value) indicated a negative correlation between isomiR composition and genotype.

SNP		isomiR		All (n = 435)			European (n = 348)			African (n = 87)		
rsID (Minor Allele)	**miRNA**	**isomiR**	**Number of bases changed**	**Tau**	***p*-value**	**FDR**	**Tau**	***p*-value**	**FDR**	**Tau**	***p*-value**	**FDR**
rs6505162 (C)	hsa-miR-423-3p	5-ext	+2	−0.37	4.9×10^{-23}	3.0×10^{-21}	−0.33	2.2×10^{-15}	9.6×10^{-14}	−0.20	0.025	0.20
rs6505162 (C)	hsa-miR-423-5p	5-trim	−2	−0.33	4.6×10^{-19}	2.4×10^{-17}	−0.30	6.5×10^{-13}	2.4×10^{-11}	−0.37	2.2×10^{-5}	9.0×10^{-4}

2.4 Discussion

The discovery of miRNA and other small non-coding RNAs has expanded our vision of the gene regulation network remarkably. The capability of a single miRNA to regulate hundreds of genes provides simultaneous control over multiple pathways. In fact, increasing evidence supports the notion that miRNAs play critical roles in diverse aspects of biological processes, and dysfunction or aberrant expression of miRNAs and their isoforms may trigger disease pathogenesis. The presence of miRNA isoforms has enriched the sequence variations within miRNAs, expanding the scope of post-transcriptional regulation of target mRNAs. The biogenesis of miRNAs and isomiRs is intricately regulated, encompassing alternative cleavage by Drosha/Dicer, which could result in base trimming and templated extension (Neilsen et al., 2012; Panzade et al., 2022). Additionally, post-transcriptional editing may contribute to the occurrence of non-templated sequence variations (Rodriguez-Galan et al., 2021). Notably, non-templated base additions have been consistently observed at a significantly higher frequency compared with sequencing errors (Tomasello et al., 2021; Wyman et al., 2011). SNPs located in a precursor or mature miRNA may influence the biogenesis, maturation, expression, or target recognition of that miRNA by altering the secondary structure of the miRNA hairpin and subsequent enzyme cleavage or transcript editing of its isomiRs. In this way, mir-SNPs may play important roles in signaling pathways that are essential to cellular homeostasis and contribute to disease progression.

In this study, we investigated isomiR expression profiles and the cis-regulation of isomiR biogenesis in human lymphoblastoid cell lines. We found that numerous SNPs were significantly associated with the frequency of 5'-end isomiRs, including base substitution, trimming, extension, and addition. Empirical evidence was observed to support the premise

that genetic variants contribute to the composition of 5'-end isomiRs by altering the sequence of precursor miRNA. Herein, we reported 95 significant associations (FDR < 0.05) between mir-SNPs and the composition of 5'-isomiRs across 435 available subjects and in a subset of either European or African populations.

The mir-SNP rs6505162 (A > C) is located in the transcribed region of precursor hsa-mir-423, but outside of the mature miRNAs. The pathologic risk of the rs6505162 polymorphism has been evaluated in a wide range of cancers and diseases, including esophageal squamous cell carcinoma, ovarian cancer, colorectal cancer, non-small-cell lung cancer, and others (Hu et al., 2008; Kontorovich et al., 2010; Xing et al., 2012; Ye et al., 2008). In breast cancer, studies reporting the association between rs6505162 and cancer risk have come to contradictory conclusions. Smith et al. showed that the CC genotype was linked to a reduced risk of breast cancer in Caucasian women (odds ratio, OR=0.50; *p*=0.03) (Smith et al., 2012), and the A allele was reported as a risk factor in the pathogenesis of breast cancer among the Egyptian population (OR=3.28, *p*=0.002; OR=2.11, *p*=0.011; for AA and CA against CC patients, respectively) (Li et al., 2015). In contrast, the CC genotype of rs6506162 was reported to be associated with an increased risk of breast cancer in Iranian women (OR=2.37; *p*=0.0023) (Pourmoshir et al., 2020). While this mir-SNP was not predicted to affect hsa-mir-423 precursor RNA secondary structure (Kontorovich et al., 2010; Zhao et al., 2015), it could in-fluence the processing efficiency and maturation of the miRNA and thereby affect breast cancer susceptibility. However, it is largely unknown how hsa-mir-423 isomiRs are in-volved in the genetic association of rs6505162 with breast cancer. In our study, we found that rs6505162 was significantly associated with hsa-miR-423-3p and hsa-miR-423-5p isomiR compositions

and that the C allele was associated with lower expression of these isomiRs. The corresponding isomiRs were found to be highly expressed in tumors compared with normal samples in TCGA-BRCA and TCGA-KIRC datasets, which suggests that the C allele may have a protective effect in tumorigenesis. This conclusion is consistent with the fact that cells with a C allele expressing the pre-miR-423 had lower proliferation than cells with the A allele (Zhao et al., 2015). Zhao et al. also reported that although rs6505162 regulated both -3p and -5p miRNAs, miR-423-3p was the only molecule promoting breast cell proliferation (Zhao et al., 2015). Because this SNP is located outside of the mature miRNA, it would not be expected to affect miR-423 binding with its targets but rather could affect the ex-pression of miR-423-3p isomiRs. Additionally, the 5'-extension isomiRs could potentially be involved in regulating the canonical hsa-miR-423-3p's target recognition, degradation, or binding to new targets implicated in breast cancer pathogenesis.

In addition, multiple complex-disease-associated SNPs identified in previous re-ports were found to be associated with isomiR compositions in human lymphoblastoid cell lines (Chen et al., 2018; Ghanbari et al., 2017; Li et al., 2015; Zhao et al., 2014). For instance, rs2273626 is located in the seed region of miR-4707 and is associated with primary open-angle glaucoma (POAG) independently of canonical miRNA expression levels (Ghanbari et al., 2017). This SNP showed a significant positive association with the composition of a 5'-extension isomiR of miR-4707-3p, implying its potential role in regulating the biological function of the canonical miRNA. Similarly, rs2168518, associated with blood pressure, triglycerides, total cholesterol, fasting glucose levels, and risk of diabetes mellitus (Li et al., 2015), was found to be associated with 5'-addition isomiRs of hsa-miR-4513 in our study.

One limitation of our study is that we used the default alignment settings for isomiRID and only considered one-base substitutions in the sequence; therefore, not all possible isomiR variations have been captured. Another limitation is that trimming the 3'-adaptor from the sRNA-seq reads made it difficult to accurately estimate the variations in 3'-isomiRs. Additionally, the associations between SNPs and isomiR compositions identified in our study are not yet supported by experimental validation, and the relationship between 5'-end base substitution isomiRs and miRNA half-lives requires further investigation in a large dataset. Despite these limitations, this study provides new insights into the genetic regulation of isomiR biogenesis in human cells and has potential implications for regulating miRNA expression and for generating new targeted therapies. In conclusion, this study investigated the cis-regulation of isomiR biogenesis in human lymphoblastoid cell lines and found significant associations between SNPs and 5'-isomiRs. Our findings, particularly the identified association between rs6505162 and isomiR alterations of hsa-miR-423-3p and hsa-miR-423-5p, shed light on the genetic aspects of breast cancer tumorigenesis. Additionally, our study revealed the prevalence of non-canonical miRNAs and allele-specific expressions of miRNAs, highlighting their roles in biological processes and the influence of genetic variants on miRNA regulation. These insights contribute to our understanding of the intricate mechanisms governing isomiR biogenesis and their implications in disease.

Chapter Three: Aberrant Intron-retention Derived Neoantigen Burden Predicts Response to Immune Checkpoint Inhibitor Therapy

3.1 Literature Review

Immune checkpoint inhibitors (ICIs) targeting CTLA-4 or PD-1/PD-L1 have been a major addition to the therapeutic landscape of a variety of tumor types and clinical settings. Since the first approval of Ipilimumab, an anti-CTLA-4 monoclonal antibody, for the treatment of metastatic melanoma in 2011, the Food and Drug Administration (FDA) has subsequently approved six more ICIs for various cancer subtypes(Jardim et al., 2021). The use of ICIs, either as monotherapy or in combination with cytotoxic therapies, has resulted in improved patient survival outcomes in a wide range of cancer subtypes(Bagchi et al., 2021; Larroquette et al., 2021). Despite these substantial advances, not all patients gain benefits from ICI agents, and many experience a unique set of immune-related adverse events (irAEs)2. As many as 70% of patients treated with ICIs encounter acute irAEs, while up to 40% of patients suffer from chronic toxicities(Johnson et al., 2022; Patrinely et al., 2021). Of particular concern are severe, potentially life-threatening side effects such as pneumonitis, myocarditis, and hypophysitis. Given the likelihood of these clinically impactful toxicities, it is imperative to better understand which patients are most likely to benefit from ICI therapy.

Considerable work has been devoted to identifying biomarkers capable of predicting response to ICI treatment. Immunohistochemical (IHC) staining for PD-L1 has been extensively evaluated in various settings but has yielded highly disparate results(Davis & Patel, 2019; Doroshow et al., 2021). This inconsistency arises from a multitude of factors, encompassing technical and biological considerations. From a

technical perspective, the type of antibody used, the cut-point defining positivity, and variability in interpretation, may impact the predictive capacity(Doroshow et al., 2021). From a biological standpoint, the importance of PD-L1 expression in treatment benefit varies by tumor type and can also depend on settings within a given cancer subtype. For example, PD-L1 positivity predicts benefit of pembrolizumab treatment for metastatic TNBC, but not for curative TNBC where pembrolizumab is equally effective regardless of PD-L1 status(Jacobson, 2022). In addition, both microsatellite instability-high (MSI-H) and tumor mutational burden-high (TMB-H) are FDA-approved biomarkers for pembrolizumab across all metastatic solid tumors. TMB-H led to the FDA approval of pembrolizumab based on data from KEYNOTE-158, which revealed an improved response rate of 29% for TMB-H patients compared to 6% for those with lower TMB(Marabelle et al., 2020). Despite an improved response rate, TMB-H did not predict significant improvements in progression- free survival (PFS) or overall survival (OS). Moreover, there is poor overlap and concordance across these biomarkers, reinforcing the variability in their predictive capacities(Wang et al., 2021).

The rationale behind utilizing TMB-H as a predictive biomarker for ICI therapy lies in its correlation with increased neoantigen expression(Moeckel et al., 2023). Tumor neoantigens are cancer-specific peptides that are absent in normal human tissue. They can be presented on the cell surface by the major histocompatibility complex (MHC) class I molecules where they are recognized by T lymphocytes, and ultimately trigger the anti-tumor immune response. A high tumor neoantigen burden is hypothesized to enhance the activity of ICIs. However, TMB quantifies somatic mutations at the DNA level, which is distal to the protein-level neoantigens. These DNA mutations may not be sufficiently

expressed in tumor cells and presented on the cell surface to elicit an anti-tumor response, which limits the accuracy of TMB as a predictor of ICI response(Katzir et al., 2022). Therefore, while TMB correlates with neoantigen burden, the relationship is not entirely congruent. Other genetic aberrancies beyond point mutations (e.g., gene fusions and dysregulated RNA splicing) can also generate neoantigens. Notably, RNA splicing-derived neoantigens are more prevalent in cancer than those acquired from point mutations (Kahles et al., 2018; Stanley & Abdel-Wahab, 2022). Intron retention is a form of RNA splicing that occurs at the more proximal mRNA level, where an intron is preserved within the mature mRNA instead of being spliced out. Aberrant intron retention can produce tumor neoantigens and it has been implicated in cancer development (Dvinge & Bradley, 2015; Smart et al., 2018). Prior studies from our group found that neoantigens resulting from aberrant intron retention are associated with survival in multiple myeloma and pancreatic cancer (Dong et al., 2021; Dong et al., 2022). The impact of aberrant intron-retention neoantigen burden (INB) on response to ICI therapy, however, is unknown. In this study, we evaluated INB as a biomarker for predicting the efficacy of ICI therapy across various metastatic tumors in an observational cohort. We hypothesized that the combination of TMB and INB would provide a more comprehensive assessment of the tumor neoantigen burden, and used together may improve the ability to predict response to ICI therapies (***Figure 3.1***).

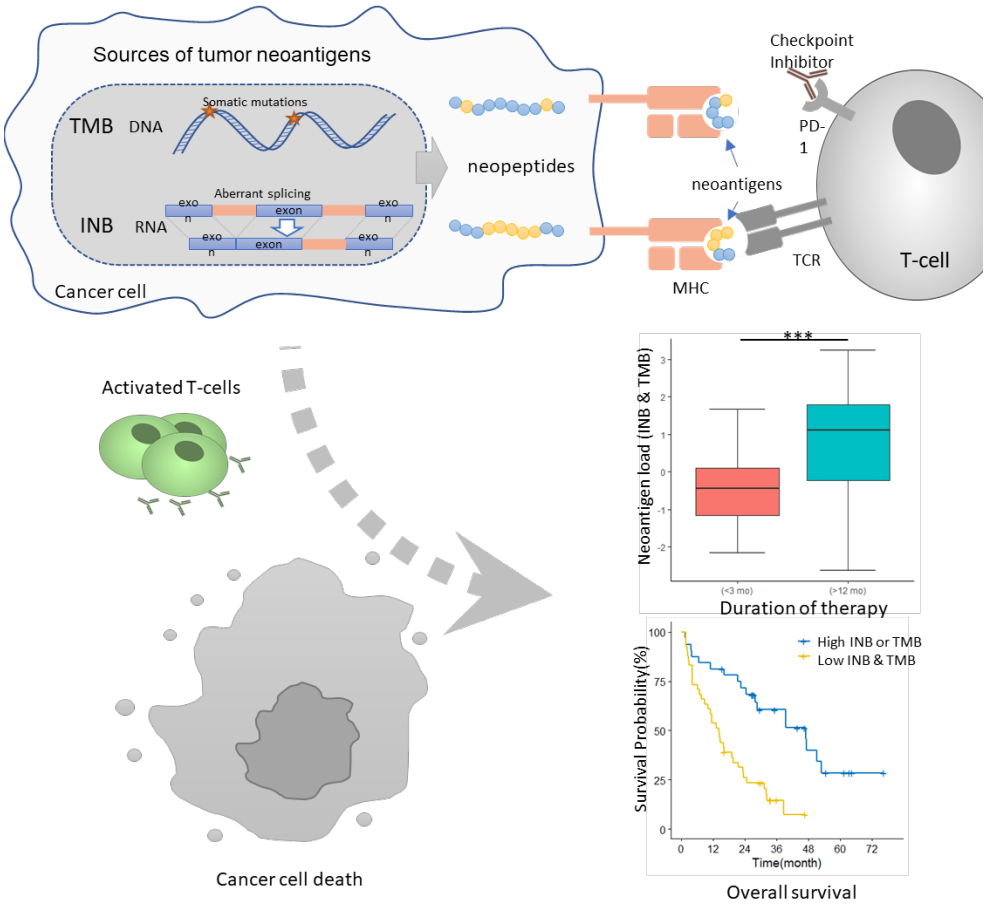


Figure 3.1. Study Schema.

3.2 Methods

3.2.1 Patient Population

Patients with advanced cancer referred to the Indiana University Health Precision Genomics (IUHPG) Program and who received ICI therapy between January 2015 and November 2020 were assessed for eligibility in this study. These patients underwent next-generation sequencing (NGS) as part of their standard of care, and recommendations for therapy were made by a multidisciplinary molecular tumor board (MTB). Treatment decisions were ultimately made by the patients' primary oncology providers. Data on therapy administration, treatment duration, and outcomes after evaluation by the MTB were obtained by chart review. TMB was assessed in genomic DNA by whole-exome sequencing (GPS Cancer®, NantHealth®). All patients included in this observational study signed informed consent for evaluation.

3.2.2 RNA Sequencing and Data Processing

Total RNA-seq libraries were prepared using the KAPA stranded RNA-seq with RiboErase kit. Aligned BAM files provided by the clinical sequencing vendor NantHealth were converted into fastq files using Samtools (version 1.15.1), and mapped to the reference genome GRCh38 using the STAR aligner (version 2.7.3a) with a two-pass mapping strategy (Dobin et al., 2013). In summary, the sequence reads were first aligned to the reference genome to identify splice junctions. The junctions from all samples were merged. The sample-specific novel junctions were derived by excluding the junctions on mitochondrial DNA, junctions with noncanonical sequences at donor and acceptor splice sites, annotated junctions previously documented, and junctions with fewer than two

uniquely mapped reads. We then created an indexed reference genome incorporating these merged junctions and used it for the second-pass alignment.

3.2.3 Prediction of Aberrant Intron-retention Neoantigen

Neoantigens derived from aberrant intron retention events were identified from the RNA-seq data as previously described (Dong et al., 2021; Dong et al., 2022). In brief, exon sets of protein-coding genes were re-annotated based on the union of exons documented in the GTF file from GENCODE.v32 database. The exon sets were used to define introns as the interval between exons. Uniquely mapped reads toward newly annotated introns and exons were quantified using HTseq (version 2.0). The identified intron retention events were filtered based on their expression levels. In addition, we excluded intron retention events that were identified in normal human tissues, which were derived from the RNA-seq data downloaded from the Gene Expression Omnibus (dataset GSE120795), using the same second-pass mapping procedure based on the reference genome indexed with the IUHPG dataset splice junctions. Patient HLA genotyping was estimated using arcasHLA (version 0.5.0). The binding affinity of the HLA allele for aberrant intron retention- derived neopeptides was assessed using NetMHCpan (version 4.1). NetMHCpan compares raw prediction scores to a set of randomly selected natural peptides to calculate the % rank, which provides robust binding metrics. Genes with aberrant intron retention-derived neopeptides whose MHC binding affinity rank in the top 2% among all peptides were counted towards neoantigen burden.

3.2.4 Statistical Analysis

Statistical analyses were performed to assess the association between INB, TMB, and clinical outcomes. The duration of therapy (DoT) was categorized as either exceptional

clinical benefit (>12 months) or no clinical benefit (< 3 months). To assess the association between demographic or clinical characteristics with DoT, we employed the Student's t-test for continuous variables and the Chi-squared test for categorical variables. The differences in INB, log-transformed TMB or their summation between patients with short and long DoT were evaluated with a t-test. Prior to the summation of INB and TMB as an indicator of the total neoantigen load, both were standardized to a mean of zero and standard deviation of one. Multivariable logistic regression models were employed to evaluate the association of INB and TMB with DoT, where long DoT was treated as an event (Y=1) and short DoT was treated as baseline (Y=0). In this model, a positive correlation coefficient or odds ratio greater than one indicated that as the independent variable increased, patients were more likely to have a longer treatment duration or a better outcome. A stepwise logistic regression, combining backward elimination and forward selection based on Akaike information criterion (AIC), was conducted to identify potential covariates (Zhang, 2016). Demographic and clinical covariates including age, gender, body mass index (BMI), Eastern Cooperative Oncology Group (ECOG) performance status, number of lines of prior therapies, tumor type and TMB were evaluated. Covariate with a p-value less than 0.05 was retained in the final regression model. We used a likelihood ratio test to compare two nested statistical models. Additionally, we applied simple linear regression to examine the relationship between INB and log transformed TMB.

Kaplan-Meier curves were utilized to estimate survival probability. Hazard ratio (HR) and its 95% confidence intervals (CIs) were assessed using a Cox proportional hazards model. Potential demographic and clinical covariates were deemed to be included in the Cox regression if their p-values were less than 0.05. Statistical analyses and data

visualization were conducted in R (version 4.2.1). Survival analyses and visualization were performed using the survival (version 3.4) and survminer (version 0.4.9) packages. Two-sided tests were used for all statistical analyses, and a p-value less than 0.05 was considered statistically significant.

We employed a leave-one-out cross-validation (LOOCV) to evaluate the performance of a predictive model. In each iteration, the model was fitted on all observations except one, which was reserved as the validation set. This process was repeated until all observations in the dataset had been used once as a validation set. The performance measures, including area under the receiver operating characteristic curve (ROC AUC) and precision-recall curve (PR AUC), obtained from the left-out observations were aggregated across all iterations to evaluate the model's overall performance.

3.3 Results

3.3.1 Patient Data

A total of 146 cancer patients who received ICI therapy were assessed for the treatment outcome. The eligible patients had metastatic solid tumors and received a minimum of two doses of ICI therapy. They also had no history of ICI treatments in the non-metastatic setting, no concurrent chemotherapy, and no therapy discontinuation within 12 months owing to drug toxicity. Patients with primary central nervous system tumors such as astrocytoma and glioblastoma were excluded, and this study exclusively focused on White patients. Out of these 146 patients met these criteria, 78 had pre-treatment RNA-seq data available, of which five patients failed HLA typing from the RNA-seq data. A final cohort of 51 patients was used for our primary study comparing therapy duration, including 31 with short DoT (< 3 months) and 20 with long DoT (> 20 months); 22

patients with DoT between 3 and 12 months were excluded in the primary comparison but included in the survival analysis (***Figure 3.2***). Patient demographics are summarized in **Table 3.1**. Information regarding the type of anti-PD-1/PD-L1 or anti-CTLA-4 inhibitors used, and the cancer types under each DoT group are outlined in **Table 3.2**.

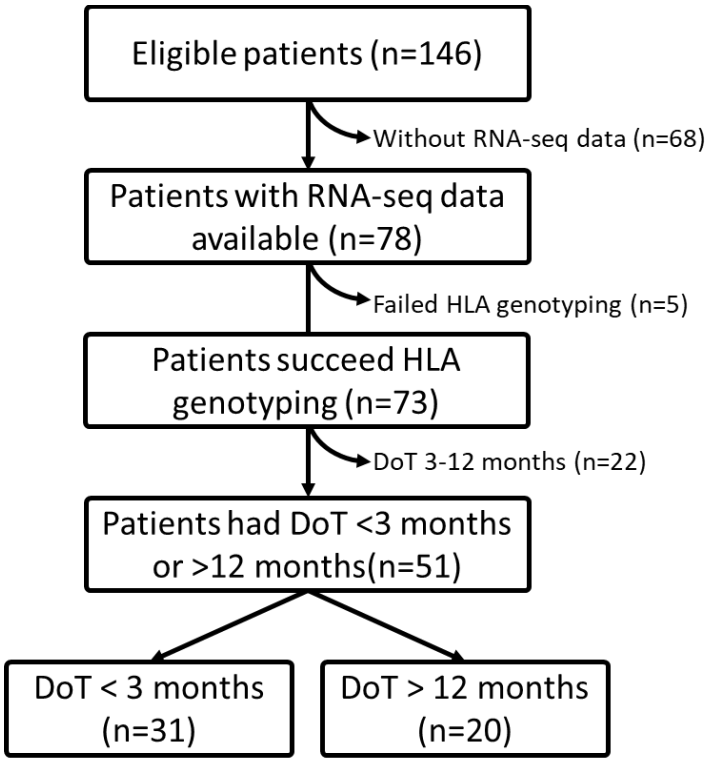


Figure 3.2. CONSORT diagram.

HLA: The human leukocyte antigen.
DoT: Duration of immune checkpoint inhibitor therapy.

Table 3.1. Demographic and clinical variables.

Characteristic	DoT < 3 mon (n=31)	DOT > 12 mon (n=20)	*p*-value
Age, years			0.58
Mean	59.1	61.1	
SD	15.9	10.0	
Race, n			NA
White	31	20	
Sex, n(%)			1
Male	20(64.5)	13(65.0)	
Female	11(35.5)	7(35.0)	
BMI			0.52
Mean	27.0	28.3	
SD	5.4	7.9	
ECOG PS, n(%)			0.21
0	10(32.3)	9(45.0)	
1	15(48.4)	10(50.0)	
2	4(12.9)	0(0)	
NA	2(6.4)	1(5.0)	
Cancer types, n(%)			0.31
head and neck	6(19.4)	3(15.0)	
melanoma	1(3.2)	7(35.0)	
non-small cell lung	5(16.1)	1(5.0)	
sarcoma	3(9.7)	2(10.0)	
bladder/urothelial	3(9.7)	1(5.0)	
renal	2(6.5)	1(5.0)	
breast	3(9.7)	0(0)	
anal	1(3.2)	1(5.0)	
esophageal	1(3.2)	1(5.0)	
skin, squamous	0(0)	1(5.0)	
thyroid	0(0)	1(5.0)	
colorectal	1(3.2)	0(0)	
hepatocellular	1(3.2)	0(0)	
neuroendocrine	1(3.2)	0(0)	
pancreas	1(3.2)	0(0)	
prostate	1(3.2)	0(0)	
unknown primary	1(3.2)	1(5.0)	

Table 3.2. Summary of immune checkpoint inhibitors received and patient cancer types

Medications	Targets	DoT < 3m (n=31)	DoT > 12mon (n=20)
Atezolizumab	PD-L1	5 (3 bladder/urothelial; 2 NSCLC)	0
Ipilimumab	CTLA-4	1(Colorectal)	0
Nivolumab	PD-1	11 (2 renal; 2 sarcoma; 2 head and neck; 1 breast; 1 hepatocellular; 1 NSCLC; 1 pancreas; 1 unknown)	6 (anal; esophageal; NSCLC; renal; skin; thyroid)
Nivolumab /Ipilimumab	PD-1 /CTLA-4	1(sarcoma)	6 (4 melanoma; 2 sarcoma)
Pembrolizumab	PD-1	13 (4 head and neck; 2 breast; 2 NSCLC; 1 prostate; 1 anal; 1 esophageal; 1 melanoma; 1 neuroendocrine)	8 (3 head and neck; 3 melanoma; 1 bladder/urothelial; 1 unknown)

DoT: Duration of ICI therapy; PD-1: programmed cell death protein 1; PD-L1: programmed cell death ligand 1; CTLA-4: cytotoxic T-lymphocyte associated protein 4; NSCLC: non-small cell lung cancer.

3.3.2 Higher INB and TMB Levels in Patients with Long Duration of ICI Therapy

The difference in INB between patients with short or long DoT was evaluated using the Student's t-test. Higher average INB was observed in pre-treatment samples from patients who had long DoT (2,041 vs. 1,753, p=0.037; ***Figure 3.3A***). Patients with long DoT also had high TMB, although the difference was not statistically significant (25.9 vs 6.4 mut/Mb; p=0.10; ***Figure 3.3B***). Additionally, as an indicator of the neoantigen load from both sources, the summation of INB and log-transformed TMB was significantly high in patients with long DoT (p=0.0045; ***Figure 3.3C***).

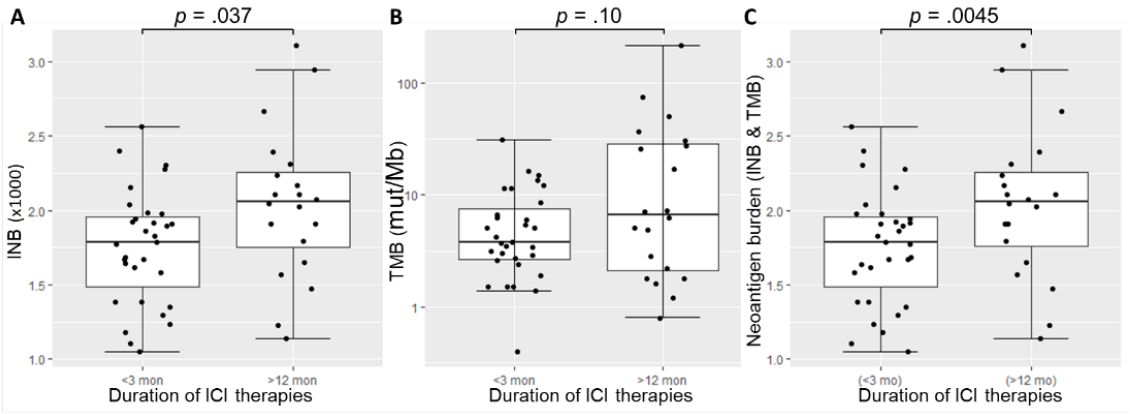


Figure 3.3. Boxplot for aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) for short (<3 months) and long (>12 months) DoT patients.

A) INB, B) log-scaled TMB, and C) the sum of standardized INB and log-transformed TMB

3.3.3 INB and TMB Were Independently Associated with ICI Therapy Duration

We next asked whether INB and TMB were correlated in pre-treatment tumor samples. Although a trend of negative correlation was observed between INB and TMB (***Figure 3.4***), an in-depth correlation analysis between the estimated INB levels and log-transformed TMB showed that these two measurements were not correlated (r=-0.14; p=0.32). This finding suggests that the INB and TMB represent independent aspects of neoantigen burden.

To gain further insight into the key factors influencing response to ICI therapies, we conducted a multivariable logistic regression analysis. Clinical and demographic covariates were selected based on a stepwise regression process. Notably, both INB and TMB emerged as significant predictors of the efficacy of the ICI therapies (p=0.019 and 0.038, respectively; **Table 3.3**). Importantly, the model that included both INB and TMB exhibited substantial improvements over models that included only one of the predictors (INB or TMB alone, likelihood ratio test p=0.0023 and 0.011, respectively). Furthermore, dichotomized INB and TMB by their 75th percentile (INB > 2,105 and TMB > 11.85 mut/Mb) retained significant associations with the DoT (p=0.028 for both INB and TMB). The distribution of INB and TMB are illustrated in ***Figure 3.5***, with the 75th percentile cutoff indicates as a vertical line.

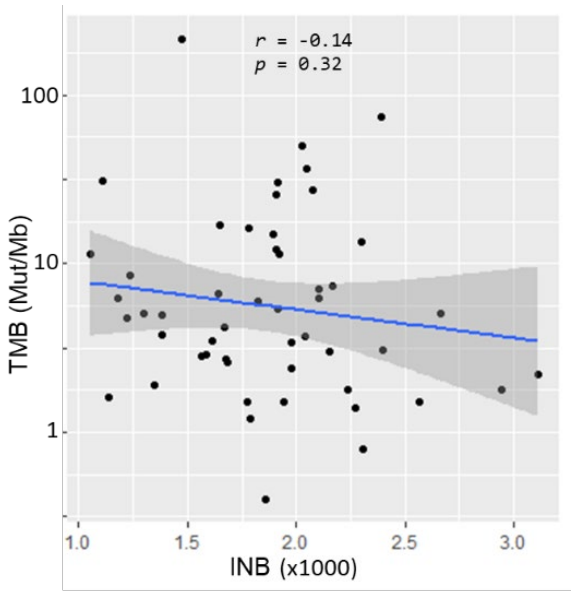


Figure 3.4. Scatter plot of aberrant intron-retention neoantigen burden (INB) and log-scaled tumor mutation burden (TMB).

The line and band indicate the fitted linear regression line and its 95% confidence intervals between intron-retention neoantigen burden (INB) and log-scaled tumor mutation burden (TMB), respectively.

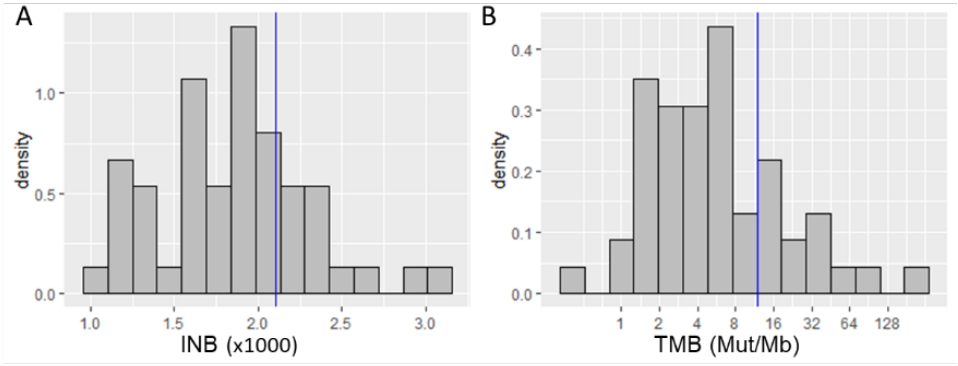


Figure 3.5. Histogram for aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB)

A) INB and B) log scaled TMB. The vertical line indicates the 75th percentile

Table 3.3. Logistic regression for duration of ICI therapy with aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) levels.

	Coefficient	SE	OR (95% CI)	p-value
INB (x1000)	1.88	0.80	6.56 (1.52 - 38.15)	0.019
TMB(Mut/Mb)	0.075	0.036	1.08 (1.02 - 1.17)	0.038

SE: standard error for the estimated regression coefficient; OR: odds ratio; CI: confidence interval.

To mitigate potential biases arising from differences in ICI drugs, we further limited our analysis to the patients treated with drugs present in both long and short DoT groups, namely Nivolumab, Nivolumab/Ipilimumab and Pembrolizumab as shown in **Table 3.4**. Six patients were excluded from this analysis since they were administered Atezolizumab and Ipilimumab, which were used only in the short DoT group. The associations of INB and TMB with DoT were still significant in the remaining 45 patients (p=0.017 and 0.049, respectively). Additionally, to account for potential confounding effects from differences in tumor types, we incorporated tumor type as a covariate in the logistic regression. INB remained significantly associated with DoT (p=0.035), while TMB did not reach statistical significance (p=0.20).

Table 3.4. Summary of immune checkpoint inhibitors received and patient cancer types

Medications	Targets	DoT < 3m (n=31)	DoT > 12mon (n=20)
Atezolizumab	PD-L1	5 (3 bladder/urothelial; 2 NSCLC)	0
Ipilimumab	CTLA-4	1(Colorectal)	0
Nivolumab	PD-1	11 (2 renal; 2 sarcoma; 2 head and neck; 1 breast; 1 hepatocellular; 1 NSCLC; 1 pancreas; 1 unknown)	6 (anal; esophageal; NSCLC; renal; skin; thyroid)
Nivolumab /Ipilimumab	PD-1 /CTLA-4	1(sarcoma)	6 (4 melanoma; 2 sarcoma)
Pembrolizumab	PD-1	13 (4 head and neck; 2 breast; 2 NSCLC; 1 prostate; 1 anal; 1 esophageal; 1 melanoma; 1 neuroendocrine)	8 (3 head and neck; 3 melanoma; 1 bladder/urothelial; 1 unknown)

DoT: Duration of ICI therapy;
PD-1: programmed cell death protein 1;
PD-L1: programmed cell death ligand 1;
CTLA-4: cytotoxic T-lymphocyte associated protein 4;
NSCLC: non-small cell lung cancer.

3.3.4 INB and TMB Predict ICI Therapy Duration in Cross Validation

To further evaluate the potential of utilizing INB and TMB as predictors of DoT on ICI therapy, we performed a leave-one-out cross-validation (LOOCV) procedure using a logistic regression model. When both INB and TMB were included as predictors, the LOOCV yielded an area under the receiver operating characteristic curve (ROC AUC) of 0.72 and precision-recall curve (PR AUC) of 0.74 (***Figure 3.6***). Importantly, this model outperformed predictions made using either INB or TMB alone (ROC AUC=0.63 and PR AUC=0.56 for INB; ROC AUC=0.52 and PR AUC=0.58 for TMB), as well as combinations of INB and TMB with demographic variables (such as age, sex, body mass index (BMI), Eastern Cooperative Oncology Group (ECOG) performance status, and prior therapies) (ROC AUC=0.57 and PR AUC=0.57).

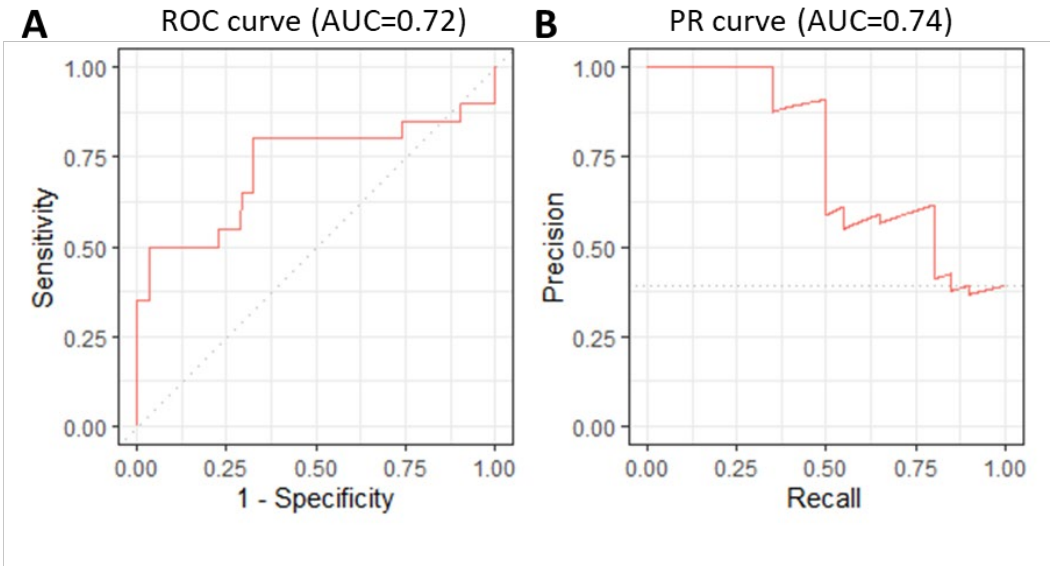


Figure 3.6. Receiver operating characteristic (ROC) and precision recall (PR) curves.

The A) ROC curve, and B) PR curve from the leave-one-out cross validation procedure with both aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB) as predictors.

3.3.5 INB and TMB Predict ICI Treatment-related Overall Survival

To gain a more comprehensive understanding of the predictive capacity of INB and TMB in the context of ICI therapy, our investigation expanded beyond the binary assessment of DoT, to incorporate their impact on patients' OS. OS times were calculated from the initiation of ICI therapy to the dates of death from any reason, with censoring occurring at six months after the conclusion of the last ICI treatment. To ensure an unbiased evaluation of the prediction power of INB and TMB on survival, we evaluated all patients (i.e., including the 22 patients with DoT between 3 and 12 months) in the survival analysis, which expanded the total sample size to 73 (***Figure 3.2***). Our analysis revealed that patients with high INB (>75th percentile, n=18) had an enhanced OS compared to those with low INB (≤75th percentile, n=55) (Hazard ratio, HR=0.45; p=0.028) (***Figure 3.7A***). Similarly, our analysis suggested that patients with high TMB based on >75th percentile exhibited superior OS compared to those with low TMB (HR=0.31, p=0.0031).

When both INB and TMB levels were considered concurrently, patients with either high INB or TMB (>75th percentile in either measurement, n=32) experienced enhanced survival outcome compared to those with low INB and TMB (≤ 75th percentile in both measurements, n=41) (HR=0.28; $p=1.1\times10^{-4}$), (***Figure 3.7B***). Survival analysis conducted with the original cohort of 51 patients consistently revealed a correlation between INB and OS (HR=0.36, p=0.024), and a strong correlation between the combination of INB and TMB with OS (HR=0.21, $p=2.1\times10^{-4}$). These findings underscore the robust predictive value of jointly considering neoantigens induced by both intron retention and tumor mutations for overall survival.

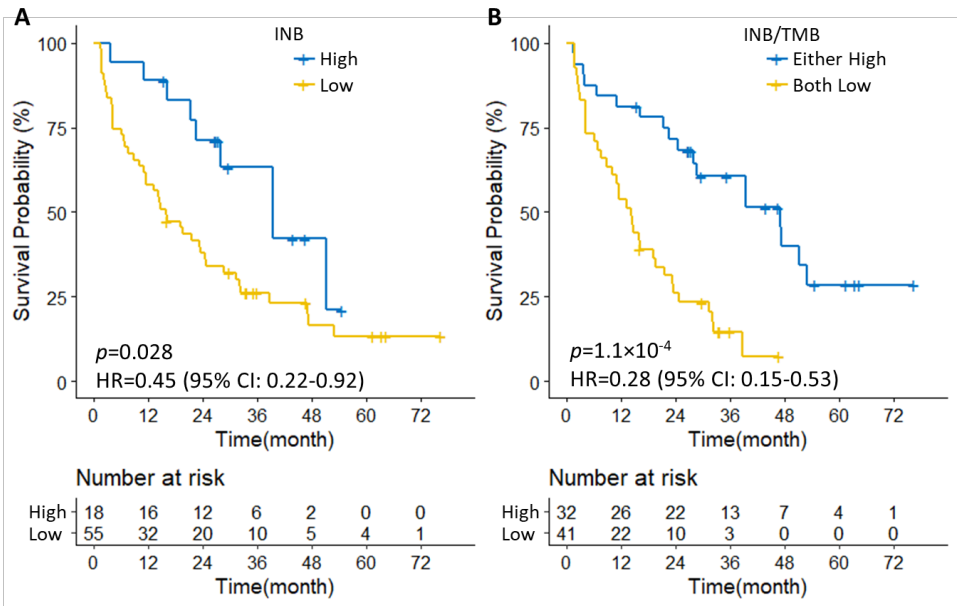


Figure 3.7. Kaplan-Meier curves for overall survival grouped by aberrant intron retention neoantigen burden (INB) and tumor mutation burden (TMB).

A) patients with high INB (> 75th percentile) vs. those with low INB (≤ 75th percentile).
B) patients with either high INB or TMB (>75th percentile in either measurement) vs. those with low INB and TMB (≤ 75th percentile in both measurements)

3.4 Discussion

The search for robust predictive biomarkers for ICI has been an ongoing challenge. Three FDA-approved biomarkers, microsatellite instability (MSI), TMB, and PD-L1 expression, have been routinely used in clinical practice for selecting patients suited for ICI treatment. Among these, PD-L1 expression, while widely studied and clinically implemented in practice, has significant limitations. MSI, the first pan-tumor FDA-approved biomarker for predicting response to ICI therapy, has impactful sensitivity when instability is identified, but its low prevalence limits the application. Many MSI-stable patients may also benefit from ICI therapy and would be left untreated if selected on MSI status alone. TMB, also designated as a pan-tumor FDA-approved biomarker for ICI therapy, exhibits modest discriminatory capacity and modest improvements in outcome. Further, there is controversy as to whether the optimal cut-point for TMB-H may better be determined based on the median baseline TMB for a given tumor type (Samstein et al., 2019). As a source of tumor neoantigens, aberrant intron retention has been implicated in the pathogenesis and therapeutic landscape of various cancers. In this report, we evaluated the predictive capacity of INB for clinical benefit from ICI therapy. We found that patients with a high INB were more likely to have experienced substantial clinical benefit on ICI therapy and had a superior OS. High TMB was also predictive, but the lack of concordance with high INB suggests that they may be complementary, rather than overlapping, predictive biomarkers.

The anti-tumor immune response relies on the recognition and binding of foreign antigens that are presented on the tumor cell surface, making neoantigens a promising biomarker for cancer immunotherapy efficacy. While the direct identification and

quantification of neoantigens through mass spectrometry (MS) is possible, the technology is not cost efficient nor high throughput (Guan et al., 2023; Q. Wang et al., 2019). Thus, surrogate markers to predict neoantigen load have become commonplace with rapid advancements in sequencing technologies and the human leukocyte antigen (HLA) binding affinity prediction approaches. Previous studies have indicated that tumor neoantigen burden based on nonsynonymous mutations is an effective biomarker for predicting response to ICI therapy (Rizvi et al., 2015; Snyder et al., 2014; Van Allen et al., 2015). However, the consideration of non-synonymous mutations alone overlooks neoantigens originating from other sources such as aberrant mRNA splicing. TMB, which evaluates point mutations at the DNA level, has been shown to correlate with neoantigen load but represents only a fraction of the genomic events that potentially contribute to the production of a neoantigen on a tumor cell. INB, which assessed neoantigen load at the RNA level, complements TMB by capturing a novel dimension of neoantigens arising from intron retention events, potentially providing additional insights into neoantigen load in cancer. The combined assessment of TMB and INB offers a more comprehensive evaluation of neoantigen burden and thus enhances predictive accuracy for ICI therapy outcomes.

A prior investigation in melanoma patients did not find a uniform difference in in intron-retention neoantigen burden among immunotherapy response groups (Smart et al., 2018). The same study did, however, highlight a case where a patient benefitted from ICI despite possessing a low mutation-based neoantigen load but a high intron retention-based neoantigen burden (Smart et al., 2017). The reasons for differences in conclusions are likely multifactorial. It is plausible that in disease types with a high abundance of

nonsynonymous mutations, such as melanoma, somatic mutation biomarkers may play a more prominent role in identifying those who will benefit. Additionally, differences in study design and analysis strategy may have contributed to the variance in findings.

A limitation of our study is the retrospective nature, conducted outside of a prospective study designed specifically to address the impact of INB on ICI therapy. The study cohort selected here, however, was well-suited to address this hypothesis as it had clinical outcomes and comprehensive sequencing performed as part of standard care. While assessing formal response rate or progression-free survival outside of a carefully controlled clinical trial is challenging, we used a clinically relevant definition of DoT as an indicator of benefit. Further, we compared extreme phenotypes, exceptional benefit vs. no clinical benefit, to amplify any biological signals of successful predictive capacity, and assessed OS as a highly reproducible and important endpoint. Another limitation was the attempt to evaluate the success of INB across tumor types. To account for this confounder, we corrected for tumor type in our multi-variate analysis. This approach, however, provided provocative data to support INB's role as a potential tumor-agnostic biomarker; similar to the current use of TMB. Additionally, our model for predicting aberrant intron-retention neoantigen burden from RNA-seq data has some limitations, including not currently account for Immunogenicity. In future studies, we plan to consider incorporating deep learning methods, such like DeepImmuno(Li et al., 2021), to improve the accuracy of predicting intron-retention neoantigen burden from RNA-seq data. Finally, the small sample size in the retrospective study limits its statistical power. To address this, we aim to validate these findings using a larger dataset. In collaboration with Immune-Oncology scientists, we are leveraging resources from larger cohort for both disease-specific and

disease-agnostic analyses. One such resource is the Oncology Research Information Exchange Network (ORIEN). ORIEN is an alliance of cancer centers across the United States that working collaboratively to accelerate cancer research and improve patient outcomes. Its extensive, clinically annotated tissue repository provides an opportunity to evaluate the performance INB and TMB in predicting patients' response to ICI therapies.

In conclusion, aberrant intron-retention neoantigen burden is a novel and promising biomarker for predicting improved response to ICI therapy. To date, no single biomarker has consistently demonstrated accuracy in predicting ICI therapy response, and improvements may be achievable through more comprehensive or complementary approaches. INB and TMB, reflecting different aspects of neoantigen formation, appear to successfully predict ICI-based therapy outcomes. This work highlights a potentially promising approach for patient stratification and personalized treatment in the evolving landscape of cancer immunotherapy.

Chapter Four: Circulating Tumor DNA in Prediction of Disease Recurrence in Triple-Negative Breast Cancer

4.1 Literature Review

Breast cancer is the most common cancer and second leading cause of cancer death among women worldwide (Roheel et al., 2023). It is a heterogeneous disease, with distinct subtypes that influence treatment options and prognosis. One particularly aggressive subtype is triple-negative breast cancer (TNBC), characterized by the lack of expression of estrogen-receptor (ER) and progesterone-receptor (PR), and the absence of human epidermal growth factor receptor 2 (HER2) amplification(Carey et al., 2010; Pal et al., 2011; Schneider et al., 2008). Although TNBC accounts for only 15-20% of breast cancer cases, it is associated with a more aggressive disease course, a higher likelihood of relapse (especially within the first three years after chemotherapy and surgery) and fatal distant metastasis, and a shorter overall survival compared to ER- and HER2-positive subtypes(Dent et al., 2007; Liedtke et al., 2023). This aggressiveness is particularly prominent in patients with residual disease after neoadjuvant chemotherapy (NAC), a standard treatment approach for early-stage TNBC patients(Cortazar et al., 2014). A pathologic complete response (pCR) after NAC is associated with an exceptional cure rate, and likely would not benefit from the addition of more systemic therapy. Conversely, approximately two-thirds of TNBC patients have residual disease. This persistent disease burden places patients at a markedly increased risk of relapse and inferior overall survival (Liedtke et al., 2023). Therefore, the “post-neoadjuvant setting” is a rich environment for exploration of therapies to improve outcomes for the high risk, and accurate identification of them is paramount for guiding personalized treatment decisions and potentially

improving outcomes through early interventions. Currently, post-NAC treatment risk assessment in TNBC relies primarily on clinical and pathological factors assessed at diagnosis or surgery, along with emerging tools like molecular subtyping(Akhouayri et al., 2022) and analysis of tumor-infiltrating lymphocytes (TILs) (Aswolinskiy et al., 2023; Leon-Ferre et al., 2024; Loi et al., 2019). Unfortunately, these tools often have limitations in accurately identifying which patients with residual disease are most likely to experience recurrence. This highlights a critical gap in TNBC management and underscores the urgent need for more precise tools for post-treatment monitoring and risk assessment.

Liquid biopsies offer a promising avenue for revolutionizing disease monitoring and recurrence prediction in cancer. The ability to non-invasively detect tumor characteristics through blood samples presents exciting potential for overcoming limitations of traditional risk assessment methods. Circulating tumor DNA (ctDNA), which contains fragments of DNA released from dying tumor cells, and circulating tumor cells (CTCs), which have detached from the primary or metastatic tumor, hold particular promise (Crowley et al., 2013; Diaz & Bardelli, 2014). Highly sensitive next-generation sequencing techniques have demonstrated their capability to detect low quantity of ctDNA in various cancer types, including breast cancer (Crowley et al., 2013; Dawson, Rosenfeld, & Caldas, 2013; Hamakawa et al., 2015; Schutz et al., 2015). Growing evidence suggests that the presence of ctDNA and CTCs in the bloodstream reflect tumor dynamics and potentially predicts disease recurrence. Notably, Chen et al. demonstrated that ctDNA in TNBC patients after NAC and surgery indicates a higher risk of rapid relapse (Chen et al., 2017). Similarly, Garcia-Murillas et al. observed a correlation between ctDNA detection and rapid recurrence in early breast cancer (Garcia-Murillas et al., 2015). Olsson et al.

showed that serial ctDNA sampling in patients with primary breast cancer reached an average lead time of 11 months before the occurrence of metastatic disease(Olsson et al., 2015). These findings were further strengthened by a later study by Garcia-Murillas et al., which demonstrated the utility of ctDNA in predicting recurrence and an inferior relapse-free survival in breast cancer (Garcia-Murillas et al., 2019).

Circulating tumor cells (CTCs) offer another valuable analyte in liquid biopsies. These cells, derived from primary or metastatic tumors, are shed into bloodstream (Racila et al., 1998). CTCs are frequently detected in both early and late-stage breast cancers, and their presence has been linked to metastatic potential and poorer patient prognosis in breast cancer (Bidard et al., 2018; Bidard et al., 2014; Cristofanilli, 2006; Smerage et al., 2014). While ctDNA generally exhibits higher detection sensitivity than CTCs in patients with metastatic breast cancer (Dawson, Tsui, et al., 2013), there are scenarios where CTCs can be isolated from the circulation in the absence of detectable levels of ctDNA. This primarily occurs when the index mutations are not covered by the ctDNA assay or due to very low concentrations/shedding of ctDNA. In a study of metastatic TNBC patients, Madic et al. found that combining ctDNA and CTC detection increased overall sensitivity compared to using either biomarker alone(Madic et al., 2015). Further, in 3 of 31 metastatic TNBC patients profiled in the study, CTCs were detected in the absence of detectable ctDNA, suggesting a population where CTCs adds sensitivity for disease detection above ctDNA.

This study aims to investigate the prognostic significance of ctDNA and CTCs in predicting disease recurrence among TNBC patients following neoadjuvant chemotherapy and surgery. We hypothesize that the presence of these circulating tumor-derived markers will be indicative of a significantly higher risk of relapse and adverse clinical outcomes,

independent of established prognostic factors. These findings have the potential to transform post-neoadjuvant treatment monitoring in TNBC. By enabling the development of more accurate risk-stratification models, these findings hold the promise of enhancing patient care through the implementation of informed, personalized interventions. Ultimately, this study aims to contribute to the optimization of treatment strategies and the improvement of clinical outcomes for TNBC patients.

4.2 Method

4.2.1 Clinical trial and sample acquisition

This study utilized samples from the clinical trial BRE12-158 (NCT02101385). BRE12-158 was a phase II randomized controlled trial investigating genomically-directed therapy for TNBC patients with significant residual disease following NAC. Patients were considered ER and PR-negative if fewer than 10% of cells stained weakly positive. HER2 was considered negative if scored 0 or 1+ by immunohistochemistry (IHC) or 2+ by IHC and associated with a fluorescence in situ hybridization (FISH) ratio of less than 2.0 or fewer than 6 copies per cell. Conducted by the Hoosier Cancer Research Network (HCRN) at 25 community and academic sites across the United States, the trial enrolled 191 TNBC patients who were treated with NAC and were found to have residual disease at the time of surgery. The residual disease was sequenced using a panel-based Clinical Laboratory Improvement Amendments (CLIA) NGS test (ParadigmDx). Sequencing results of residual disease from the surgical specimen, along with the patient's medical and treatment history, were reviewed by a molecular tumor board at Indiana University. Patients with actionable alterations were randomized to either genomically-directed therapy (Arm A) or treatment of physician's choice (Arm B). Patients without actionable

targets were assigned to Arm B (***Figure 4.1***). The primary objective of the trial was to compare 2-year DFS in randomly allocated patients in arm A versus arm B. Following the trial protocol, all patients were registered on trial no less than 14 days (2 weeks) and no later than 84 days (12 weeks) from completion of their last loco-regional therapy (either surgery or radiation). Plasma samples were collected either prior to treatment on Cycle 1 Day1 for Arm A, or at first routine visit for Arm B (***Figure 4.1***), and ctDNA was isolated using the Qiagen circulating nucleic acids kit. All patients were consented for the trial and correlative studies and the protocol approved by the Indiana University Institutional Review Board (IRB) and the IRBs of all participating sites. Patient clinical demographics, including self-reported race and ethnicity, treatment history, and tumor characteristics were collected at enrollment.

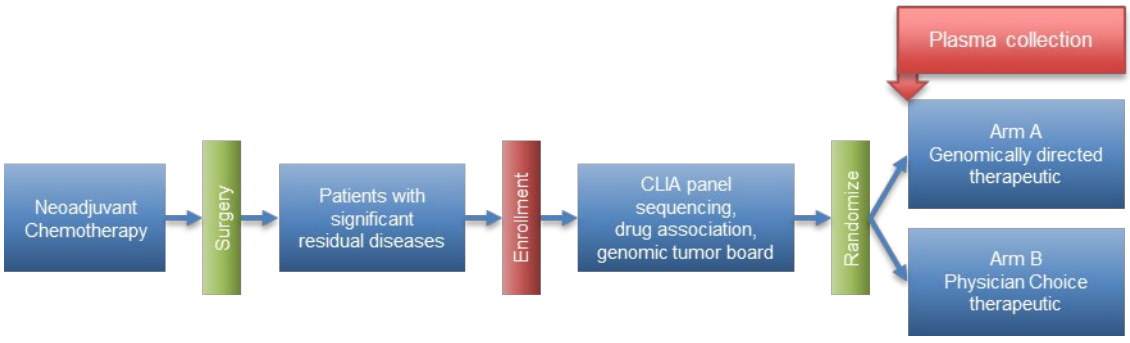


Figure 4.1 BRE12-158 Trial Schema.

A Phase II Randomized Controlled Trial of Genomically Directed Therapy After Preoperative Chemotherapy in Patients with Triple Negative Breast Cancer

4.2.2 Sequencing and data processing

A minimum of 20 mL of peripheral whole blood was collected in K3 EDTA tubes and 20-100 ng cell-free DNA (cfDNA) was extracted from plasma for subsequent analysis. Samples underwent comprehensive genomic profiling using hybridization-captured, adaptor ligation-based libraries. Tests cover either 62 (all coding exons of 27 genes, select exons of 34 genes, FoundationACT®, n = 95) or 70 (all coding exons of 35 genes, select exons of 35 genes, FoundationOne Liquid®, n = 105) cancer-related genes, plus selected introns from 6 (FoundationACT™) or 7 (FoundationOne Liquid™) genes frequently rearranged in cancer. Profiling examined base substitutions, small insertions or deletions, focal copy number alterations (amplifications), and rearrangements as previously described (Clark, Chung et al. 2018). Liquid biopsy testing was conducted in a Clinical Laboratory Improvement Amendments–certified, College of American Pathologists–accredited reference laboratory (Foundation Medicine, Cambridge, MA). ctDNA content was estimated using maximum somatic allele frequency (MSAF), as described in Li et al. 2019 (Li et al., 2019). In brief, this method calculates the fraction of tumor-derived ctDNA versus that of total cfDNA by leveraging the allele fraction of non-germline alterations, as well as aneuploidy in the specimen's copy number profile. To ensure consistency, only genomic content shared by both FoundationACT™ and FoundationOne Liquid™ was analyzed. Samples failing QC were excluded.

4.2.3 Detection of somatic mutations from ctDNA sequencing

ctDNA samples were sequenced for commonly mutated oncogenes. Mutations detected in ctDNA were carefully filtered to ensure the highest likelihood of somatic in nature. Patients were considered ctDNA-positive based on the following criteria: (1)

Detection of point mutation(s) or insertion/deletion(s) present at a variant allele frequency(VAF) <40% or >60% (ultra-low allele frequency indels were excluded as these were most likely false-positive findings); (2) Point mutations or indels with VAF between 40-60% were excluded if present in ClinVar or dbSNP databases as known germline variants.; (3) Detection of copy number amplifications or deletions; or (4) Detection of a genomic rearrangement event. All patients who harbored somatic mutations also had a positive MSAF.

4.2.4 CTC isolation and enumeration

Blood samples for CTC analysis were collected in CellSave Preservative tubes (Menarini Silicon Biosystems) and shipped at room temperature to Purdue University. One 10mL tube was collected per datapoint. Samples containing at least 4mL of blood were processed within 3 days from collection using an established method (Chang et al., 2015). Briefly, anti-EpCAM antibodies conjugated to magnetic beads were incubated with blood for 1-2 hours. Subsequently, the blood is processed through a microfluidic device designed for rare cell analysis. Nucleated cells (<1000) captured on the device were further stained using anti-pan-cytokeratin-FITC, anti-CD45-PE, and DAPI as previously described(Chang et al., 2015; Chang et al., 2014). Nucleated cells captured on the device staining positive for FITC and DAPI and negative for CD45 were called CTCs. The output for each datapoint was recorded as the absolute number of CTCs detected.

4.2.5 Statistical analysis

Survival time was measured from the date of randomization to the date of event of interest (or death from any cause), or the date of censoring (loss to follow-up). Kaplan-Meier curve was used to estimate the survival probabilities and to compare survival

outcomes between patient subgroups. Multivariate Cox proportional-hazards regression was used to assess the association between ctDNA/CTC status and survival while control for potential confounding variables. Demographic or clinical variables considered in the model included age, race, tumor size at surgery, residual cancer burden (RCB) classifications (II or III), histological grade, pathologic tumor stage and pathological nodes stage at surgery. A stepwise variable selection procedure based on the Akaike information criterion (AIC) was employed within Cox proportional-hazards model to identify significant covariates for the final regression analyses. Only covariates with a p-value < 0.05 were retained in the final model. Chi-square test was used to investigate the association between categorical variables like CTC and ctDNA status. A significance level of p-value < 0.05 was used throughout the study.

4.3 Results

4.3.1 Patient demographics and sample selection

196 patients were enrolled on the BRE12-158 clinical trial with 142 evaluable patients for our pre-planned correlative analysis of ctDNA and 123 evaluable patients for our correlative analysis of CTCs. 113 patients had both ctDNA and CTC data available. Details of patient selection included in this study are outlined in the Consolidated Standards for Reporting Trials (CONSORT) diagram (***Figure 4.2***). Patient characteristics, including: demographics, neoadjuvant therapy, stage, grade, tumor size, lymph node involvement, and Residual Cancer Burden (RCB) classification are detailed in Table 4.1.

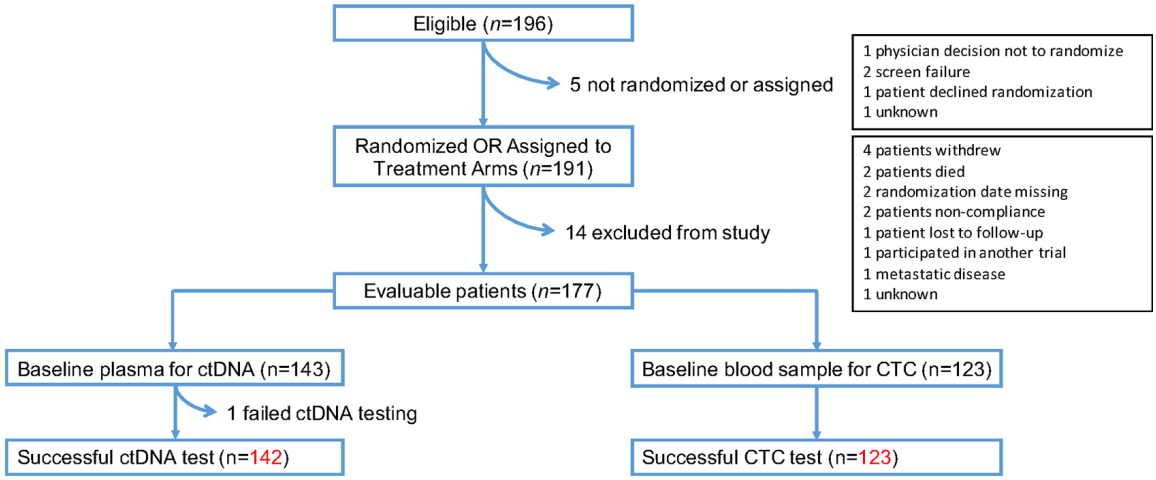


Figure 4.2 CONSORT diagram.

CONSORT: Consolidated Standards of Reporting Trials

Table 4.1. Clinical and pathological characteristics of patients analyzed for plasma ctDNA in the BRE12-158 trial.

Characteristics	All patients analyzed for plasma ctDNA, No. (%) (n=196)	ctDNA - and + patients No. (%) (n=142)		CTC - and + patients No. (%) (n=123)	
		Negative	Positive	Negative	Positive
Age group (in years)		***P*=.07**		***P*=.15**	
≤ 45	71 (36)	18(35)	28(31)	20(27)	21(42)
46-60	87 (44)	28(54)	36(40)	37(51)	18(36)
61-75	36 (18)	6(12)	25(28)	16(22)	10(20)
≥ 76	2 (1)	0(0)	1(1)	0(0)	1(2)
Race		***P*=.19**		***P*=.36**	
White	141 (72)	39(75)	75(83)	55(75)	41(82)
Black	50 (26)	12(23)	13(14)	17(23)	7(14)
Asian	3 (2)	0(0)	2(2)	1(1)	1(2)
Unknown	2 (1)	1(2)	0(0)	0(0)	1(2)
Ethnicity		***P*=.72**		***P*=.67**	
Hispanic or Latino	17 (9)	5(10)	9(10)	9(12)	5(10)
Non-Hispanic or Latino	175 (89)	45(87)	80(89)	63(86)	43(86)
Unknown	4 (2)	2(4)	1(1)	1(1)	2(4)
Arms		***P*=.63**		***P*=.09**	
Arm A	70 (36)	20(38)	37(41)	28(38)	21(42)
Arm B (Randomized)	73 (37)	22(42)	31(34)	34(47)	15(30)
Arm B (Assigned)	48 (24)	10(19)	22(24)	10(14)	14(28)
Not Randomized or Assigned	5 (3)	0(0)	0(0)	1(1)	0(0)
Pathologic T Stage at Surgery		***P*=.66**		***P*=.27**	
TX	11 (6)	2(4)	6(7)	6(8)	1(2)
T1	75 (38)	22(42)	35(39)	33(45)	17(34)
T2	79 (40)	20(38)	35(39)	26(36)	22(44)
T3	21 (11)	7(13)	8(9)	6(8)	8(16)
T4	10 (5)	1(2)	6(7)	2(3)	2(4)
Pathologic N Stage at Surgery		***P*=.21**		***P*=.29**	
NX	3 (2)	0(0)	1(1)	1(1)	0(0)
N0	104 (53)	34(65)	43(48)	43(59)	27(54)
N1	50 (26)	10(19)	29(32)	20(27)	16(32)
N2	21 (11)	5(10)	7(8)	2(3)	5(10)
N3	18 (9)	3(6)	10(11)	7(10)	2(4)
Pathologic M Stage at Surgery		***P*=.04**		***P*=.70**	

MX	63 (32)	12(23)	37(41)	22(30)	17(34)
M0	132 (67)	40(77)	53(59)	51(70)	33(66)
M1	1 (1)	0(0)	0(0)	0(0)	0(0)
Anatomic Stage		***P*=.35**		***P*=.88**	
I	45 (23)	16(31)	19(21)	21(29)	12(24)
II	100 (51)	26(50)	46(51)	39(53)	28(56)
III	50 (26)	10(19)	25(28)	13(18)	10(20)
Unknown	1 (1)	0(0)	0(0)	0(0)	0(0)
Histologic Grade		***P*=.71**		***P*=0.52**	
Grade 1	1 (1)	0(0)	1(1)	1(1)	0(0)
Grade 2	22 (11)	7(13)	12(13)	10(14)	8(16)
Grade 3	168 (86)	45(87)	74(82)	59(81)	42(84)
Unknown	5 (3)	0(0)	3(3)	3(4)	0(0)
Lymph Node Involvement		***P*=.12**		***P*=.58**	
Yes	90 (46)	19(37)	46(51)	29(40)	23(46)
No	104 (53)	33(63)	44(49)	44(60)	27(54)
Unknown	2 (1)	0(0)	0(0)	0(0)	0(0)
Tumor Size at Surgery		***P*=.84**		***P*=.40**	
≤ 2 cm	88 (45)	23(44)	43(48)	37(51)	20(40)
2 ~ 5 cm	80 (41)	21(40)	36(40)	25(34)	23(46)
> 5 cm	27 (14)	8(15)	11(12)	11(15)	7(14)
Unknown	1 (1)	0(0)	0(0)	0(0)	0(0)
RCB Classification		***P*=.52**		***P*=.68**	
II	86 (44)	21(40)	43(48)	39(53)	23(46)
III	37 (19)	9(17)	18(20)	10(14)	9(18)
Unknown	73 (37)	22(42)	29(32)	24(33)	18(36)

4.3.2 Association of ctDNA With Survival Outcomes

ctDNA positivity was consistent across both treatment arms, detected in 65% of patients (37 of 57) in arm A and 62% of patients (53 of 85) in arm B. The median follow-up time was 17.2 months (range, 0.1-58.3 months). Detection of ctDNA was significantly associated with survival outcomes in the study cohort. ctDNA-positive patients exhibited short DDFS compared to ctDNA-negative patients (hazard ratio [HR] 2.99; 95% CI 1.38-6.48; *p*=0 .006) (***Figure 4.3A***). At 24 months, the DDFS probability was 56% for ctDNA-positive patients compared with 81% for ctDNA-negative patients. Similarly, detection of ctDNA was significantly associated with an inferior DFS (HR 2.67; 95% CI 1.28-5.57; *p* =0 .009) (***Figure 4.3B***). At 24 months, the DFS probability was 50% for ctDNA-positive patients compared with 76% for ctDNA-negative patients. Last, detection of ctDNA was significantly associated with an inferior OS (HR 4.16; 95% CI 1.66-10.42; *p*=0 .002) (***Figure 4.3C***). At 24 months, the OS probability was 57% for ctDNA-positive patients compared with 80% for ctDNA-negative patients.

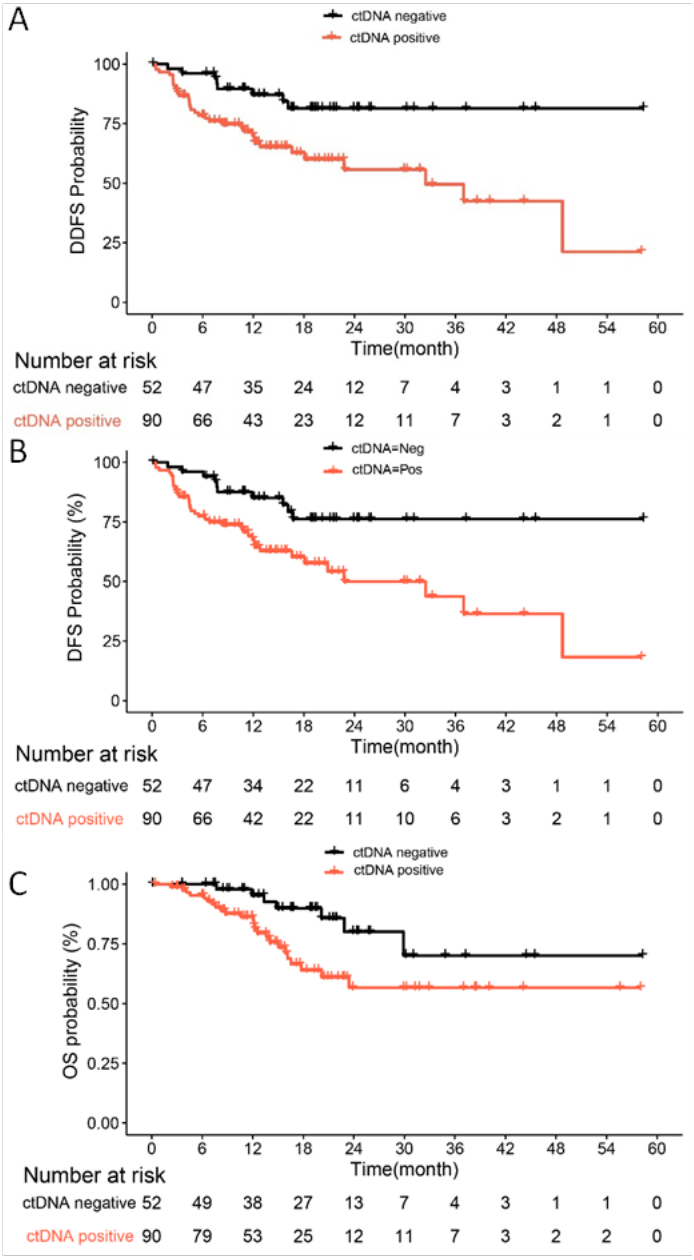


Figure 4.3. Association of the presence of ctDNA with survival outcomes.

(A) Distant disease-free survival (DDFS); median DDFS:32.5 months vs. Not Reached; HR=2.99, 95%CI:1.38-6.48; *p*=0.0055. At 24 months, the DDFS probability was 56% in ctDNA-positive patients as compared to 81% in ctDNA-negative patients. **(B)** Disease-free survival (DFS); median DFS:22.8 months vs. Not Reached; HR=2.67, 95%CI:1.28-5.57; *p*=0.0089). At 24 months, the DFS probability was 50% in ctDNA-positive patients as compared to 76% in ctDNA-negative patients. **(C)** Overall survival (OS); median OS: Not Reached vs. Not Reached; HR=4.16, 95%CI:1.66-10.42; *p*=0.0024. At 24 months, the OS probability was 57% in ctDNA-positive patients as compared to 80% in ctDNA-negative patients.

4.3.3 Association of the Combination of CTCs and ctDNA With Clinical Outcomes

Circulating tumor cell positivity was detected in 43% of patients (21 of 49) in arm A and 39% of patients (29 of 74) in arm B. Although patients who were CTC positive had inferior survival outcomes, the differences did not reach statistical significance (***Figure 4.4***). However, increasing CTC count was significantly associated with inferior DDFS (HR, 1.07; 95% CI, 1.01-1.13; $p = 0.02$), DFS (HR, 1.11; 95% CI, 1.03-1.19; $p = 0.004$), and OS (HR, 1.09; 95% CI, 1.02-1.17; $p = 0.01$). This highlights the importance of CTC quantity as a potential predictor of disease burden and patient outcomes, and circulating tumor cells may provide additional information about the presence of minimal residual disease (MRD). Notably, in 112 patients with both ctDNA and CTC data, we did not find a significant association between CTC positivity (defined as ≥1 CTC detected) and ctDNA positivity ($p = 0.19$). Importantly, some patients had detectable ctDNA without detectable CTCs and vice-versa. By combining ctDNA and CTC analysis, the sensitivity for recurrence detection increased from 79% (23 of 29) with ctDNA alone and 62% (18 of 29) with CTC alone to 90% (26 of 29).

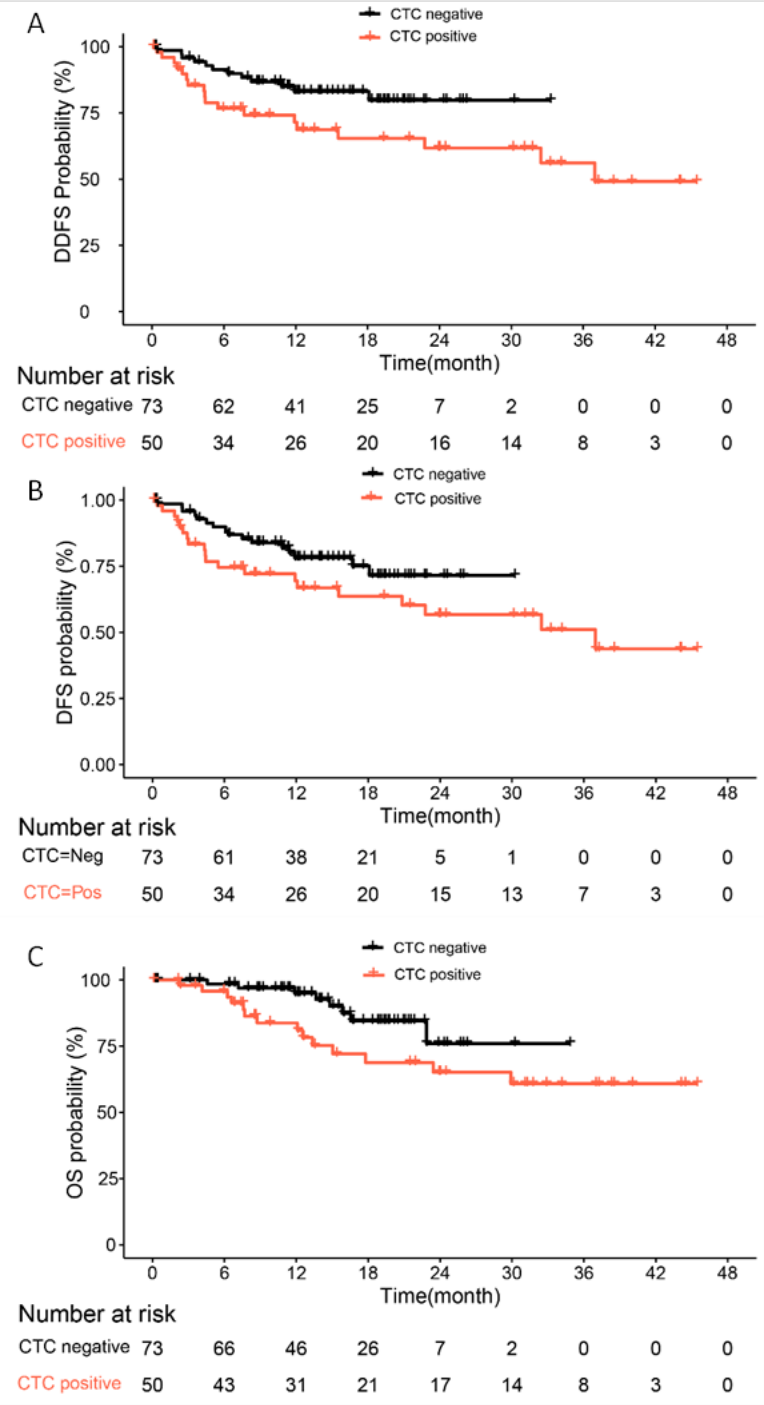


Figure 4.4. Association of the presence of CTCs with survival outcomes.

(A) distant disease-free survival (DDFS); HR=1.97, p=0.077; **(B)** disease free survival (DFS); HR=1.58, p=0.20; and **(C)** overall survival (OS); HR=2.18, p=0.10.

To gain further insights, we stratified patients based on their ctDNA and CTC status into four groups: (1) ctDNA positive / CTC positive, (2) ctDNA positive / CTC negative, (3) ctDNA negative / CTC positive, and (4) ctDNA negative / CTC negative. The DDFS curves demonstrated a stepwise gradation in which patients positive for both ctDNA and CTCs had inferior DDFS compared with those positive for either marker alone, and patients negative for both markers had the best outcomes (***Figure 4.5A***). Importantly, patients who were ctDNA positive / CTC positive exhibited significantly inferior DDFS compared to ctDNA-negative / CTC-negative (HR 5.29; 95% CI 1.50-18.62; $p = 0.009$). At 24 months, the DDFS probability was 52% for ctDNA-positive / CTC-positive patients compared to 89% in the ctDNA-negative / CTC-negative group. Similar trends were observed in DFS (HR 3.15; 95% CI 1.07-9.27; $p = 0.04$) (***Figure 4.5B***) and OS (HR 8.60; 95% CI 1.78-41.47; $p = 0.007$) (***Figure 4.5C***) among patients who were ctDNA positive / CTC positive compared with those who were ctDNA negative / CTC negative. Interestingly, patients who were ctDNA-positive/CTC-negative or ctDNA-negative/CTC-positive had comparable risks of recurrence, suggesting similar prognostic value for either biomarker detected individually.

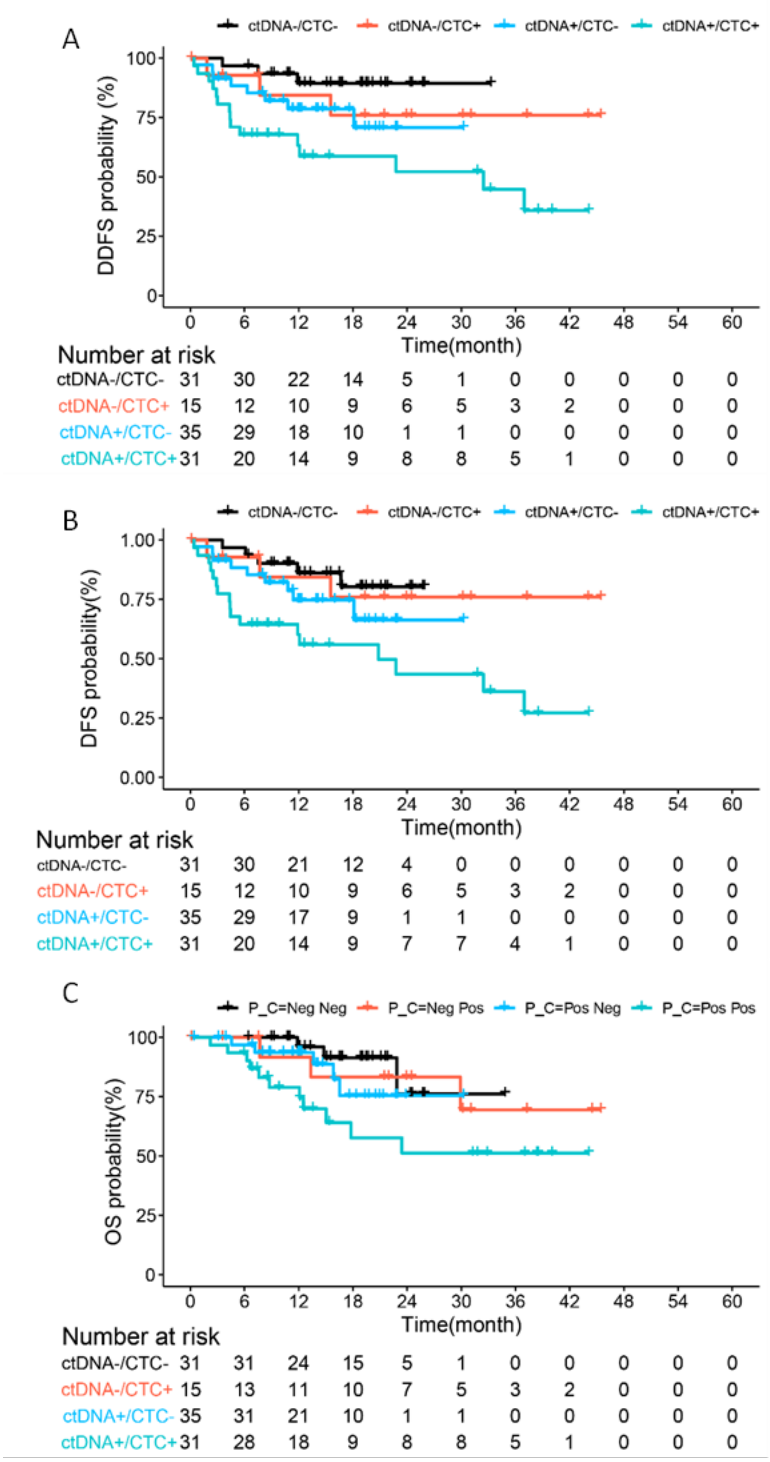


Figure 4.5. Association of presence of ctDNA and CTCs with survival outcomes.

(A) distant disease-free survival (DDFS); **(B)** disease-free survival (DFS); and **(C)** overall survival (OS).

4.4 Discussion

The detection of ctDNA in cancer patients has gained significant attention since its discovery by Leon et al. in 1977 (Leon et al., 1977). Recent advancements in next-generation sequencing have propelled ctDNA-based analysis. While ctDNA analysis has shown value in tracking metastatic breast cancer progression (Dawson, Tsui, et al., 2013), its use in detecting MRD remains an emerging area (Chen et al., 2017; Garcia-Murillas et al., 2015). This study adds substantially to the growing body of evidence by revealing the prognostic significance of ctDNA for detecting MRD in high-risk TNBC patients after NAC. This represents one of the largest datasets to date within the context of a pre-planned secondary analysis of a prospective randomized clinical trial.

The results indicate that TNBC patients with residual disease following NAC can be further risk-stratified based on the presence of MRD as determined by ctDNA. Patients without detectable ctDNA exhibited a significant improvement in DFS, DDFS, and OS, with striking effect size that remains highly significant even after accounting for multiple clinical variables. While DFS was the primary endpoint of the parent trial, DDFS is more biologically relevant as a surrogate for OS. Importantly, the ability to assess ctDNA at a single time point after surgery is highly clinically relevant, as it coincides with crucial post-neoadjuvant therapy decisions.

While ctDNA alone proved to be a robust prognostic marker, we also investigated the potential added value of CTC detection. Though sensitivity was not markedly improved by the combined analysis, the absence of both ctDNA and CTCs was associated with superior relapse-free survival. This suggests that a subset of patients with clinically

quiescent MRD might be missed by current ctDNA assays, emphasizing an area for future assay refinement.

Certain limitations of this work must be acknowledged. Potential interactions with different post-neoadjuvant therapies, while balanced across the study arms, should be considered in future investigations. Another limitation is the lack of a comprehensive mutation panel, which might have led to missed ctDNA positivity detection. The relatively short duration of follow-up is also a concern, although the early relapse pattern in this population minimizes this issue. Despite these early findings, we expect to see even further separation with additional follow-up. While relapses in the ctDNA positive population will only increase with time, it will be important to determine whether those patients in the ctDNA negative arm continue to remain disease free.

The urgent need for improvements in TNBC outcomes in post-neoadjuvant setting is a crucial area for clinical trials, especially for subgroup of patients with residual tumor after standard pre-operative therapy. Our findings support using MRD status as a major stratification variable in this setting. Additionally, the ability to sequence ctDNA broadly opens the door to identifying not only ultra-high-risk patients but also actionable drug targets. This concept will form the basis of our successor trial to BRE12-158, where ctDNA-positive patients will receive targeted therapy tailored to their plasma NGS results. Of equal importance, the ctDNA-negative/CTC-negative subgroup may represent patients who would not benefit from additional therapy, making this an ideal population for novel de-escalation strategies. While we advise against routine clinical use of ctDNA for relapse monitoring or treatment decisions outside of clinical trials, our findings, combined with the existing literature, underscore its potential for transforming risk stratification in TNBC

trials. The future of precision oncology undoubtedly lies with tools like ctDNA and their promise to revolutionize how we manage this aggressive breast cancer subtype.

References

Akhouayri, L., Ostano, P., Mello-Grand, M., Gregnanin, I., Crivelli, F., Laurora, S., Liscia, D., Leone, F., Santoro, A., Mule, A., Guarino, D., Maggiore, C., Carlino, A., Magno, S., Scatolini, M., Di Leone, A., Masetti, R., & Chiorino, G. (2022). Identification of a minimum number of genes to predict triple-negative breast cancer subgroups from gene expression profiles. *Hum Genomics*, *16*(1), 70. https://doi.org/10.1186/s40246-022-00436-6

Amsel, D., Vilcinskas, A., & Billion, A. (2017). Evaluation of high-throughput isomiR identification tools: illuminating the early isomiRome of Tribolium castaneum. *BMC Bioinformatics*, *18*(1), 359. https://doi.org/10.1186/s12859-017-1772-z

Ardekani, A. M., & Naeini, M. M. (2010). The Role of MicroRNAs in Human Diseases. *Avicenna J Med Biotechnol*, *2*(4), 161-179. https://www.ncbi.nlm.nih.gov/pubmed/23407304

Aswolinskiy, W., Munari, E., Horlings, H. M., Mulder, L., Bogina, G., Sanders, J., Liu, Y. H., van den Belt-Dusebout, A. W., Tessier, L., Balkenhol, M., Stegeman, M., Hoven, J., Wesseling, J., van der Laak, J., Lips, E. H., & Ciompi, F. (2023). PROACTING: predicting pathological complete response to neoadjuvant chemotherapy in breast cancer from routine diagnostic histopathology biopsies with deep learning. *Breast Cancer Res*, *25*(1), 142. https://doi.org/10.1186/s13058-023-01726-0

Bagchi, S., Yuan, R., & Engleman, E. G. (2021). Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. *Annu Rev Pathol*, *16*, 223-249. https://doi.org/10.1146/annurev-pathol-042020-042741

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Bhardwaj, A., Singh, H., Trinidad, C. M., Albarracin, C. T., Hunt, K. K., & Bedrosian, I. (2018). The isomiR-140-3p-regulated mevalonic acid pathway as a potential target for prevention of triple negative breast cancer. *Breast Cancer Res*, *20*(1), 150. https://doi.org/10.1186/s13058-018-1074-z

Bidard, F. C., Michiels, S., Riethdorf, S., Mueller, V., Esserman, L. J., Lucci, A., Naume, B., Horiguchi, J., Gisbert-Criado, R., Sleijfer, S., Toi, M., Garcia-Saenz, J. A., Hartkopf, A., Generali, D., Rothe, F., Smerage, J., Muinelo-Romay, L., Stebbing, J., Viens, P., . . . Pantel, K. (2018). Circulating Tumor Cells in Breast Cancer Patients Treated by Neoadjuvant Chemotherapy: A Meta-analysis. *J Natl Cancer Inst*, *110*(6), 560-567. https://doi.org/10.1093/jnci/djy018

Bidard, F. C., Peeters, D. J., Fehm, T., Nole, F., Gisbert-Criado, R., Mavroudis, D., Grisanti, S., Generali, D., Garcia-Saenz, J. A., Stebbing, J., Caldas, C., Gazzaniga, P., Manso, L., Zamarchi, R., de Lascoiti, A. F., De Mattos-Arruda, L., Ignatiadis, M., Lebofsky, R., van Laere, S. J., . . . Michiels, S. (2014). Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. *Lancet Oncol*, *15*(4), 406-414. https://doi.org/10.1016/S1470-2045(14)70069-5

Borel, C., Deutsch, S., Letourneau, A., Migliavacca, E., Montgomery, S. B., Dimas, A. S., Vejnar, C. E., Attar, H., Gagnebin, M., Gehrig, C., Falconnet, E., Dupre, Y., Dermitzakis, E. T., & Antonarakis, S. E. (2011). Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. *Genome Res*, *21*(1), 68-73. https://doi.org/10.1101/gr.109371.110

Budach, S., Heinig, M., & Marsico, A. (2016). Principles of microRNA Regulation Revealed Through Modeling microRNA Expression Quantitative Trait Loci. *Genetics*, *203*(4), 1629-1640. https://doi.org/10.1534/genetics.116.187153

Carey, L., Winer, E., Viale, G., Cameron, D., & Gianni, L. (2010). Triple-negative breast cancer: disease entity or title of convenience? *Nat Rev Clin Oncol*, *7*(12), 683-692. https://doi.org/10.1038/nrclinonc.2010.154

Chang, C. L., Huang, W., Jalal, S. I., Chan, B. D., Mahmood, A., Shahda, S., O'Neil, B. H., Matei, D. E., & Savran, C. A. (2015). Circulating tumor cell detection using a parallel flow micro-aperture chip system. *Lab Chip*, *15*(7), 1677-1688. https://doi.org/10.1039/c5lc00100e

Chang, C. L., Jalal, S. I., Huang, W. F., Mahmood, A., Matei, D. E., & Savran, C. A. (2014). High-Throughput Immunomagnetic Cell Detection Using a Microaperture Chip System. *Ieee Sensors Journal*, *14*(9), 3008-3013. https://doi.org/10.1109/Jsen.2014.2321167

Chen, Y., Du, M., Chen, W., Zhu, L., Wu, C., Zhang, Z., Wang, M., Chu, H., Gu, D., & Chen, J. (2018). Polymorphism rs2682818 in miR-618 is associated with colorectal cancer susceptibility in a Han Chinese population. *Cancer Med*, *7*(4), 1194-1200. https://doi.org/10.1002/cam4.1409

Chen, Y. H., Hancock, B. A., Solzak, J. P., Brinza, D., Scafe, C., Miller, K. D., & Radovich, M. (2017). Next-generation sequencing of circulating tumor DNA to predict recurrence in triple-negative breast cancer patients with residual disease after neoadjuvant chemotherapy. *NPJ Breast Cancer*, *3*, 24. https://doi.org/10.1038/s41523-017-0028-4

Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J. P., Wolmark, N., Bonnefoi, H., Cameron, D., Gianni, L., Valagussa, P., Swain, S. M., Prowell, T., Loibl, S., Wickerham, D. L., Bogaerts, J., Baselga, J., Perou, C., Blumenthal, G., Blohmer, J., . . . von Minckwitz, G. (2014). Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*, *384*(9938), 164-172. https://doi.org/10.1016/S0140-6736(13)62422-8

Cristofanilli, M. (2006). Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *Semin Oncol*, *33*(3 Suppl 9), S9-14. https://doi.org/10.1053/j.seminoncol.2006.03.016

Crowley, E., Di Nicolantonio, F., Loupakis, F., & Bardelli, A. (2013). Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol*, *10*(8), 472-484. https://doi.org/10.1038/nrclinonc.2013.110

Davis, A. A., & Patel, V. G. (2019). The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. *J Immunother Cancer*, *7*(1), 278. https://doi.org/10.1186/s40425-019-0768-9

Dawson, S. J., Rosenfeld, N., & Caldas, C. (2013). Circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med*, *369*(1), 93-94. https://doi.org/10.1056/NEJMc1306040

Dawson, S. J., Tsui, D. W., Murtaza, M., Biggs, H., Rueda, O. M., Chin, S. F., Dunning, M. J., Gale, D., Forshew, T., Mahler-Araujo, B., Rajan, S., Humphray, S., Becq, J., Halsall, D., Wallis, M., Bentley, D., Caldas, C., & Rosenfeld, N. (2013). Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med*, *368*(13), 1199-1209. https://doi.org/10.1056/NEJMoa1213261

de Oliveira, L. F., Christoff, A. P., & Margis, R. (2013). isomiRID: a framework to identify microRNA isoforms. *Bioinformatics*, *29*(20), 2521-2523. https://doi.org/10.1093/bioinformatics/btt424

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207-3212. https://doi.org/10.1093/bioinformatics/btp579

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P., & Narod, S. A. (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res*, *13*(15 Pt 1), 4429-4434. https://doi.org/10.1158/1078-0432.CCR-06-3045

Desvignes, T., Batzel, P., Berezikov, E., Eilbeck, K., Eppig, J. T., McAndrews, M. S., Singer, A., & Postlethwait, J. H. (2015). miRNA Nomenclature: A View Incorporating Genetic Origins, Biosynthetic Pathways, and Sequence Variants. *Trends Genet*, *31*(11), 613-626. https://doi.org/10.1016/j.tig.2015.09.002

Diaz, L. A., Jr., & Bardelli, A. (2014). Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol*, *32*(6), 579-586. https://doi.org/10.1200/JCO.2012.45.2011

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Dong, C., Cesarano, A., Bombaci, G., Reiter, J. L., Yu, C. Y., Wang, Y., Jiang, Z., Zaid, M. A., Huang, K., Lu, X., Walker, B. A., Perna, F., & Liu, Y. (2021). Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene*, *40*(42), 6130-6138. https://doi.org/10.1038/s41388-021-02005-y

Dong, C., Reiter, J. L., Dong, E., Wang, Y., Lee, K. P., Lu, X., & Liu, Y. (2022). Intron-Retention Neoantigen Load Predicts Favorable Prognosis in Pancreatic Cancer. *JCO Clin Cancer Inform*, *6*, e2100124. https://doi.org/10.1200/CCI.21.00124

Doroshow, D. B., Bhalla, S., Beasley, M. B., Sholl, L. M., Kerr, K. M., Gnjatic, S., Wistuba, II, Rimm, D. L., Tsao, M. S., & Hirsch, F. R. (2021). PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nat Rev Clin Oncol*, *18*(6), 345-362. https://doi.org/10.1038/s41571-021-00473-5

Dvinge, H., & Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome Med*, *7*(1), 45. https://doi.org/10.1186/s13073-015-0168-9

Dweep, H., Sticht, C., Pandey, P., & Gretz, N. (2011). miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform*, *44*(5), 839-847. https://doi.org/10.1016/j.jbi.2011.05.002

Faraoni, I., Antonetti, F. R., Cardone, J., & Bonmassar, E. (2009). miR-155 gene: a typical multifunctional microRNA. *Biochim Biophys Acta*, *1792*(6), 497-505. https://doi.org/10.1016/j.bbadis.2009.02.013

Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, *19*(1), 92-105. https://doi.org/10.1101/gr.082701.108

Garcia-Murillas, I., Chopra, N., Comino-Mendez, I., Beaney, M., Tovey, H., Cutts, R. J., Swift, C., Kriplani, D., Afentakis, M., Hrebien, S., Walsh-Crestani, G., Barry, P., Johnston, S. R. D., Ring, A., Bliss, J., Russell, S., Evans, A., Skene, A., Wheatley, D., . . . Turner, N. C. (2019). Assessment of Molecular Relapse Detection in Early-Stage Breast Cancer. *JAMA Oncol*, *5*(10), 1473-1478. https://doi.org/10.1001/jamaoncol.2019.1838

Garcia-Murillas, I., Schiavon, G., Weigelt, B., Ng, C., Hrebien, S., Cutts, R. J., Cheang, M., Osin, P., Nerurkar, A., Kozarewa, I., Garrido, J. A., Dowsett, M., Reis-Filho, J. S., Smith, I. E., & Turner, N. C. (2015). Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med*, *7*(302), 302ra133. https://doi.org/10.1126/scitranslmed.aab0021

Ghanbari, M., Iglesias, A. I., Springelkamp, H., van Duijn, C. M., Ikram, M. A., Dehghan, A., Erkeland, S. J., Klaver, C. C. W., Meester-Smoor, M. A., & International Glaucoma Genetics, C. (2017). A Genome-Wide Scan for MicroRNA-Related Genetic Variants Associated With Primary Open-Angle Glaucoma. *Invest Ophthalmol Vis Sci*, *58*(12), 5368-5377. https://doi.org/10.1167/iovs.17-22410

Guan, H., Wu, Y., Li, L. U., Yang, Y., Qiu, S., Zhao, Z., Chu, X., He, J., Chen, Z., Zhang, Y., Ding, H., Pan, J., & Pan, Y. (2023). Tumor neoantigens: Novel strategies for application of cancer immunotherapy. *Oncol Res*, *31*(4), 437-448. https://doi.org/10.32604/or.2023.029924

Hamakawa, T., Kukita, Y., Kurokawa, Y., Miyazaki, Y., Takahashi, T., Yamasaki, M., Miyata, H., Nakajima, K., Taniguchi, K., Takiguchi, S., Mori, M., Doki, Y., & Kato, K. (2015). Monitoring gastric cancer progression with circulating tumour DNA. *Br J Cancer*, *112*(2), 352-356. https://doi.org/10.1038/bjc.2014.609

Hanna, J., Hossain, G. S., & Kocerha, J. (2019). The Potential for microRNA Therapeutics and Clinical Research. *Front Genet*, *10*, 478. https://doi.org/10.3389/fgene.2019.00478

Hu, Z., Chen, J., Tian, T., Zhou, X., Gu, H., Xu, L., Zeng, Y., Miao, R., Jin, G., Ma, H., Chen, Y., & Shen, H. (2008). Genetic variants of miRNA sequences and non-small cell lung cancer survival. *J Clin Invest*, *118*(7), 2600-2608. https://doi.org/10.1172/JCI34934

Huan, T., Rong, J., Liu, C., Zhang, X., Tanriverdi, K., Joehanes, R., Chen, B. H., Murabito, J. M., Yao, C., Courchesne, P., Munson, P. J., O'Donnell, C. J., Cox, N., Johnson, A. D., Larson, M. G., Levy, D., & Freedman, J. E. (2015). Genome-wide identification of microRNA expression quantitative trait loci. *Nat Commun*, *6*, 6601. https://doi.org/10.1038/ncomms7601

Humphreys, D. T., Hynes, C. J., Patel, H. R., Wei, G. H., Cannon, L., Fatkin, D., Suter, C. M., Clancy, J. L., & Preiss, T. (2012). Complexity of murine cardiomyocyte miRNA biogenesis, sequence variant expression and function. *PLoS One*, *7*(2), e30933. https://doi.org/10.1371/journal.pone.0030933

Hwang, H. W., & Mendell, J. T. (2006). MicroRNAs in cell proliferation, cell death, and tumorigenesis. *Br J Cancer*, *94*(6), 776-780. https://doi.org/10.1038/sj.bjc.6603023

Jacobson, A. (2022). Pembrolizumab Improves Outcomes in Early-Stage and Locally Advanced or Metastatic Triple-Negative Breast Cancer. *Oncologist*, *27*(Suppl 1), S17-S18. https://doi.org/10.1093/oncolo/oyac014

Jardim, D. L., Goodman, A., de Melo Gagliato, D., & Kurzrock, R. (2021). The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer Cell*, *39*(2), 154-173. https://doi.org/10.1016/j.ccell.2020.10.001

Johnson, D. B., Nebhan, C. A., Moslehi, J. J., & Balko, J. M. (2022). Immune-checkpoint inhibitors: long-term implications of toxicity. *Nat Rev Clin Oncol*, *19*(4), 254-267. https://doi.org/10.1038/s41571-022-00600-w

Juvvuna, P. K., Khandelia, P., Lee, L. M., & Makeyev, E. V. (2012). Argonaute identity defines the length of mature mammalian microRNAs. *Nucleic Acids Res*, *40*(14), 6808-6820. https://doi.org/10.1093/nar/gks293

Kahles, A., Lehmann, K. V., Toussaint, N. C., Huser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research, N., & Ratsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, *34*(2), 211-224 e216. https://doi.org/10.1016/j.ccell.2018.07.001

Kai, Z. S., & Pasquinelli, A. E. (2010). MicroRNA assassins: factors that regulate the disappearance of miRNAs. *Nat Struct Mol Biol*, *17*(1), 5-10. https://doi.org/10.1038/nsmb.1762

Katzir, R., Rudberg, N., & Yizhak, K. (2022). Estimating tumor mutational burden from RNA-sequencing without a matched-normal sample. *Nat Commun*, *13*(1), 3092. https://doi.org/10.1038/s41467-022-30753-2

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81-93.

Kontorovich, T., Levy, A., Korostishevsky, M., Nir, U., & Friedman, E. (2010). Single nucleotide polymorphisms in miRNA binding sites and miRNA genes as breast/ovarian cancer risk modifiers in Jewish high-risk women. *Int J Cancer*, *127*(3), 589-597. https://doi.org/10.1002/ijc.25065

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res*, *47*(D1), D155-D162. https://doi.org/10.1093/nar/gky1141

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., . . . Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506-511. https://doi.org/10.1038/nature12531

Larroquette, M., Domblides, C., Lefort, F., Lasserre, M., Quivy, A., Sionneau, B., Bertolaso, P., Gross-Goupil, M., Ravaud, A., & Daste, A. (2021). Combining

immune checkpoint inhibitors with chemotherapy in advanced solid tumours: A review. *Eur J Cancer*, *158*, 47-62. https://doi.org/10.1016/j.ejca.2021.09.013

Lee, L. W., Zhang, S., Etheridge, A., Ma, L., Martin, D., Galas, D., & Wang, K. (2010). Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, *16*(11), 2170-2180. https://doi.org/10.1261/rna.2225110

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843-854. https://doi.org/10.1016/0092-8674(93)90529-y

Leon-Ferre, R. A., Jonas, S. F., Salgado, R., Loi, S., de Jong, V., Carter, J. M., Nielsen, T. O., Leung, S., Riaz, N., Chia, S., Jules-Clement, G., Curigliano, G., Criscitiello, C., Cockenpot, V., Lambertini, M., Suman, V. J., Linderholm, B., Martens, J. W. M., van Deurzen, C. H. M., . . . International Immuno-Oncology Biomarker Working, G. (2024). Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer. *JAMA*, *331*(13), 1135-1144. https://doi.org/10.1001/jama.2024.3056

Leon, S. A., Shapiro, B., Sklaroff, D. M., & Yaros, M. J. (1977). Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res*, *37*(3), 646-650. https://www.ncbi.nlm.nih.gov/pubmed/837366

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, *120*(1), 15-20. https://doi.org/10.1016/j.cell.2004.12.035

Li, G., Iyer, B., Prasath, V. B. S., Ni, Y., & Salomonis, N. (2021). DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Brief Bioinform*, *22*(6). https://doi.org/10.1093/bib/bbab160

Li, G., Pavlick, D., Chung, J. H., Bauer, T., Tan, B. A., Peguero, J., Ward, P., Kallab, A., Bufill, J., Hoffman, A., Sadiq, A., Edenfield, J., He, J., Cooke, M., Hughes, J., Forcier, B., Nahas, M., Stephens, P., Ali, S. M., . . . Gregg, J. P. (2019). Genomic profiling of cell-free circulating tumor DNA in patients with colorectal cancer and its fidelity to the genomics of the tumor biopsy. *J Gastrointest Oncol*, *10*(5), 831-840. https://doi.org/10.21037/jgo.2019.05.05

Li, Q., Chen, L., Chen, D., Wu, X., & Chen, M. (2015). Influence of microRNA-related polymorphisms on clinical outcomes in coronary artery disease. *Am J Transl Res*, *7*(2), 393-400. https://www.ncbi.nlm.nih.gov/pubmed/25901206

Li, Y., & Kowdley, K. V. (2012). MicroRNAs in common human diseases. *Genomics Proteomics Bioinformatics*, *10*(5), 246-253. https://doi.org/10.1016/j.gpb.2012.07.005

Liedtke, C., Mazouni, C., Hess, K. R., Andre, F., Tordai, A., Mejia, J. A., Symmans, W. F., Gonzalez-Angulo, A. M., Hennessy, B., Green, M., Cristofanilli, M., Hortobagyi, G. N., & Pusztai, L. (2023). Response to Neoadjuvant Therapy and Long-Term Survival in Patients With Triple-Negative Breast Cancer. *J Clin Oncol*, *41*(10), 1809-1815. https://doi.org/10.1200/JCO.22.02572

Loi, S., Drubay, D., Adams, S., Pruneri, G., Francis, P. A., Lacroix-Triki, M., Joensuu, H., Dieci, M. V., Badve, S., Demaria, S., Gray, R., Munzone, E., Lemonnier, J., Sotiriou, C., Piccart, M. J., Kellokumpu-Lehtinen, P. L., Vingiani, A., Gray, K., Andre, F., . . . Michiels, S. (2019). Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. *J Clin Oncol*, *37*(7), 559-569. https://doi.org/10.1200/JCO.18.01010

Madic, J., Kiialainen, A., Bidard, F. C., Birzele, F., Ramey, G., Leroy, Q., Rio Frio, T., Vaucher, I., Raynal, V., Bernard, V., Lermine, A., Clausen, I., Giroud, N., Schmucki, R., Milder, M., Horn, C., Spleiss, O., Lantz, O., Stern, M. H., . . . Lebofsky, R. (2015). Circulating tumor DNA and circulating tumor cells in metastatic triple negative breast cancer patients. *Int J Cancer*, *136*(9), 2158-2165. https://doi.org/10.1002/ijc.29265

Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., Chung, H. C., Kindler, H. L., Lopez-Martin, J. A., Miller, W. H., Jr., Italiano, A., Kao, S., Piha-Paul, S. A., Delord, J. P., McWilliams, R. R., Fabrizio, D. A., Aurora-Garg, D., Xu, L., Jin, F., . . . Bang, Y. J. (2020). Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol*, *21*(10), 1353-1365. https://doi.org/10.1016/S1470-2045(20)30445-9

Moeckel, C., Bakhl, K., Georgakopoulos-Soares, I., & Zaravinos, A. (2023). The Efficacy of Tumor Mutation Burden as a Biomarker of Response to Immune Checkpoint Inhibitors. *Int J Mol Sci*, *24*(7). https://doi.org/10.3390/ijms24076710

Neilsen, C. T., Goodall, G. J., & Bracken, C. P. (2012). IsomiRs--the overlooked repertoire in the dynamic microRNAome. *Trends Genet*, *28*(11), 544-549. https://doi.org/10.1016/j.tig.2012.07.005

Olsson, E., Winter, C., George, A., Chen, Y., Howlin, J., Tang, M. H., Dahlgren, M., Schulz, R., Grabau, D., van Westen, D., Ferno, M., Ingvar, C., Rose, C., Bendahl, P. O., Ryden, L., Borg, A., Gruvberger-Saal, S. K., Jernstrom, H., & Saal, L. H. (2015). Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. *EMBO Mol Med*, *7*(8), 1034-1047. https://doi.org/10.15252/emmm.201404913

Pal, S. K., Childs, B. H., & Pegram, M. (2011). Triple negative breast cancer: unmet medical needs. *Breast Cancer Res Treat*, *125*(3), 627-636. https://doi.org/10.1007/s10549-010-1293-1

Panzade, G., Li, L., Hebbar, S., Veksler-Lublinsky, I., & Zinovyeva, A. (2022). Global profiling and annotation of templated isomiRs dynamics across Caenorhabditis elegans development. *RNA Biol*, *19*(1), 928-942. https://doi.org/10.1080/15476286.2022.2099646

Patrinely, J. R., Jr., Johnson, R., Lawless, A. R., Bhave, P., Sawyers, A., Dimitrova, M., Yeoh, H. L., Palmeri, M., Ye, F., Fan, R., Davis, E. J., Rapisuwon, S., Long, G. V., Haydon, A., Osman, I., Mehnert, J. M., Carlino, M. S., Sullivan, R. J., Menzies, A. M., & Johnson, D. B. (2021). Chronic Immune-Related Adverse Events Following Adjuvant Anti-PD-1 Therapy for High-risk Resected Melanoma. *JAMA Oncol*, *7*(5), 744-748. https://doi.org/10.1001/jamaoncol.2021.0051

Pourmoshir, N., Motalleb, G. H., & Vallian, S. (2020). hsa-miR-423 rs6505162 Is Associated with The Increased Risk of Breast Cancer in Isfahan Central Province of Iran. *Cell J*, *22*(Suppl 1), 110-116. https://doi.org/10.22074/cellj.2020.7011

Racila, E., Euhus, D., Weiss, A. J., Rao, C., McConnell, J., Terstappen, L. W., & Uhr, J. W. (1998). Detection and characterization of carcinoma cells in the blood. *Proc Natl Acad Sci U S A*, *95*(8), 4589-4594. https://doi.org/10.1073/pnas.95.8.4589

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., . . . Chan, T. A. (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*, *348*(6230), 124-128. https://doi.org/10.1126/science.aaa1348

Rodriguez-Galan, A., Dosil, S. G., Gomez, M. J., Fernandez-Delgado, I., Fernandez-Messina, L., Sanchez-Cabo, F., & Sanchez-Madrid, F. (2021). MiRNA post-transcriptional modification dynamics in T cell activation. *iScience*, *24*(6), 102530. https://doi.org/10.1016/j.isci.2021.102530

Roheel, A., Khan, A., Anwar, F., Akbar, Z., Akhtar, M. F., Imran Khan, M., Sohail, M. F., & Ahmad, R. (2023). Global epidemiology of breast cancer based on risk factors: a systematic review. *Front Oncol*, *13*, 1240098. https://doi.org/10.3389/fonc.2023.1240098

Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., Russo, P., Rosenberg, J., Iyer, G., Bochner, B. H., . . . Morris, L. G. T. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet*, *51*(2), 202-206. https://doi.org/10.1038/s41588-018-0312-8

Schneider, B. P., Winer, E. P., Foulkes, W. D., Garber, J., Perou, C. M., Richardson, A., Sledge, G. W., & Carey, L. A. (2008). Triple-negative breast cancer: risk factors to potential targets. *Clin Cancer Res*, *14*(24), 8010-8018. https://doi.org/10.1158/1078-0432.CCR-08-1208

Schutz, E., Akbari, M. R., Beck, J., Urnovitz, H., Zhang, W. W., Bornemann-Kolatzki, K., Mitchell, W. M., Nam, R. K., & Narod, S. A. (2015). Chromosomal instability in cell-free DNA is a serum biomarker for prostate cancer. *Clin Chem*, *61*(1), 239-248. https://doi.org/10.1373/clinchem.2014.226571

Smart, A. C., Margolis, C., Miao, D., Liu, D., Park, J., He, M. X., Reardon, B., Mullane, S., Schilling, B., Garraway, L. A., Schadendorf, D., & Van Allen, E. M. (2017). Intron retention as a novel source of tumor neoantigens associated with response to checkpoint inhibitor therapy[abstract]. AACR 2017, Washington, DC. Philadelphia (PA).

Smart, A. C., Margolis, C. A., Pimentel, H., He, M. X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K. K., & Van Allen, E. M. (2018). Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol*, *36*(11), 1056-1058. https://doi.org/10.1038/nbt.4239

Smerage, J. B., Barlow, W. E., Hortobagyi, G. N., Winer, E. P., Leyland-Jones, B., Srkalovic, G., Tejwani, S., Schott, A. F., O'Rourke, M. A., Lew, D. L., Doyle, G. V., Gralow, J. R., Livingston, R. B., & Hayes, D. F. (2014). Circulating tumor cells and response to chemotherapy in metastatic breast cancer: SWOG S0500. *J Clin Oncol*, *32*(31), 3483-3489. https://doi.org/10.1200/JCO.2014.56.2561

Smith, R. A., Jedlinski, D. J., Gabrovska, P. N., Weinstein, S. R., Haupt, L., & Griffiths, L. R. (2012). A genetic variant located in miR-423 is associated with reduced breast cancer risk. *Cancer Genomics Proteomics*, *9*(3), 115-118. https://www.ncbi.nlm.nih.gov/pubmed/22593246

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., Hollmann, T. J., Bruggeman, C., Kannan, K., Li, Y., Elipenahli, C., Liu, C., Harbison, C. T., Wang, L., Ribas, A., . . . Chan, T. A. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med*, *371*(23), 2189-2199. https://doi.org/10.1056/NEJMoa1406498

Stanley, R. F., & Abdel-Wahab, O. (2022). Dysregulation and therapeutic targeting of RNA splicing in cancer. *Nat Cancer*, *3*(5), 536-546. https://doi.org/10.1038/s43018-022-00384-z

Starega-Roslan, J., Krol, J., Koscianska, E., Kozlowski, P., Szlachcic, W. J., Sobczak, K., & Krzyzosiak, W. J. (2011). Structural basis of microRNA length variety. *Nucleic Acids Res*, *39*(1), 257-268. https://doi.org/10.1093/nar/gkq727

Telonis, A. G., Magee, R., Loher, P., Chervoneva, I., Londin, E., & Rigoutsos, I. (2017). Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res*, *45*(6), 2973-2985. https://doi.org/10.1093/nar/gkx082

Tomasello, L., Distefano, R., Nigita, G., & Croce, C. M. (2021). The MicroRNA Family Gets Wider: The IsomiRs Classification and Role. *Front Cell Dev Biol*, *9*, 668648. https://doi.org/10.3389/fcell.2021.668648

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M. H. G., Goldinger, S. M., Utikal, J., Hassel, J. C., Weide, B., Kaehler, K. C., Loquai, C., Mohr, P., Gutzmer, R., Dummer, R., Gabriel, S., . . . Garraway, L. A. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, *350*(6257), 207-211. https://doi.org/10.1126/science.aad0095

van der Kwast, R., Woudenberg, T., Quax, P. H. A., & Nossent, A. Y. (2020). MicroRNA-411 and Its 5'-IsomiR Have Distinct Targets and Functions and Are Differentially Regulated in the Vasculature under Ischemia. *Mol Ther*, *28*(1), 157-170. https://doi.org/10.1016/j.ymthe.2019.10.002

Wang, J., Mei, J., & Ren, G. (2019). Plant microRNAs: Biogenesis, Homeostasis, and Degradation. *Front Plant Sci*, *10*, 360. https://doi.org/10.3389/fpls.2019.00360

Wang, Q., Douglass, J., Hwang, M. S., Hsiue, E. H., Mog, B. J., Zhang, M., Papadopoulos, N., Kinzler, K. W., Zhou, S., & Vogelstein, B. (2019). Direct Detection and Quantification of Neoantigens. *Cancer Immunol Res*, *7*(11), 1748-1754. https://doi.org/10.1158/2326-6066.CIR-19-0107

Wang, Y., Tong, Z., Zhang, W., Zhang, W., Buzdin, A., Mu, X., Yan, Q., Zhao, X., Chang, H. H., Duhon, M., Zhou, X., Zhao, G., Chen, H., & Li, X. (2021). FDA-Approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Front Oncol*, *11*, 683419. https://doi.org/10.3389/fonc.2021.683419

Woldemariam, N. T., Agafonov, O., Hoyheim, B., Houston, R. D., Taggart, J. B., & Andreassen, R. (2019). Expanding the miRNA Repertoire in Atlantic Salmon; Discovery of IsomiRs and miRNAs Highly Expressed in Different Tissues and Developmental Stages. *Cells*, *8*(1). https://doi.org/10.3390/cells8010042

Wyman, S. K., Knouf, E. C., Parkin, R. K., Fritz, B. R., Lin, D. W., Dennis, L. M., Krouse, M. A., Webster, P. J., & Tewari, M. (2011). Post-transcriptional generation of

miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res*, *21*(9), 1450-1461. https://doi.org/10.1101/gr.118059.110

Xing, J., Wan, S., Zhou, F., Qu, F., Li, B., Myers, R. E., Fu, X., Palazzo, J. P., He, X., Chen, Z., & Yang, H. (2012). Genetic polymorphisms in pre-microRNA genes as prognostic markers of colorectal cancer. *Cancer Epidemiol Biomarkers Prev*, *21*(1), 217-227. https://doi.org/10.1158/1055-9965.EPI-11-0624

Ye, Y., Wang, K. K., Gu, J., Yang, H., Lin, J., Ajani, J. A., & Wu, X. (2008). Genetic variations in microRNA-related genes are novel susceptibility loci for esophageal cancer risk. *Cancer Prev Res (Phila)*, *1*(6), 460-469. https://doi.org/10.1158/1940-6207.CAPR-08-0135

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Ann Transl Med*, *4*(7), 136. https://doi.org/10.21037/atm.2016.03.35

Zhao, H., Gao, A., Zhang, Z., Tian, R., Luo, A., Li, M., Zhao, D., Fu, L., Fu, L., Dong, J. T., & Zhu, Z. (2015). Genetic analysis and preliminary function study of miR-423 in breast cancer. *Tumour Biol*, *36*(6), 4763-4771. https://doi.org/10.1007/s13277-015-3126-7

Zhao, Y., Wei, Q., Hu, L., Chen, F., Hu, Z., Heist, R. S., Su, L., Amos, C. I., Shen, H., & Christiani, D. C. (2014). Polymorphisms in MicroRNAs are associated with survival in non-small cell lung cancer. *Cancer Epidemiol Biomarkers Prev*, *23*(11), 2503-2511. https://doi.org/10.1158/1055-9965.EPI-14-0389

Zhou, L., Lim, M. Y. T., Kaur, P., Saj, A., Bortolamiol-Becet, D., Gopal, V., Tolwinski, N., Tucker-Kellogg, G., & Okamura, K. (2018). Importance of miRNA stability and alternative primary miRNA isoforms in gene regulation during Drosophila development. *Elife*, *7*. https://doi.org/10.7554/eLife.38389

Curriculum Vitae

Guanglong Jiang

EDUCATION

2017.9-2024.7	Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis Ph.D., Bioinformatics
2013.8-2015.12	Department of Mathematical Sciences, Purdue University, Indianapolis M.S., Applied Statistics
2001.9-2004.6	Biology departments, Capital Normal University, China M.S., Genetics;
1997.9-2001.7	Biology departments, Henan Normal University, China B.S., Biology;

WORK & RESEARCH EXPERIENCE

2018.2-now	Bioinformatician Center for Computational Biology and Bioinformatics (CCBB) Department of Medical and Molecular Genetics Indiana University School of Medicine
2010.10-2017.10	Research Associate Center for Computational Biology and Bioinformatics (CCBB) Department of Medical and Molecular Genetic Indiana University School of Medicine
2009.2-2010.10	Visiting Research Scholar Scientific Programing and bioinformatics research. Center for Computational Biology and Bioinformatics (CCBB) and School of Informatics, Indiana University-Purdue University Indianapolis (IUPUI)

PUBLICATIONS

1. **Jiang, G**., Reiter, JL., Shen F., Dong C., …, Schneider BP., Liu, Y. Aberrant intron-retention derived neoantigen burden predicts response to immune checkpoint inhibitor therapy., (submitted)
2. **Jiang, G**., Reiter, J. L., Dong, C., Wang, Y., Fang, F., Jiang, Z., & Liu, Y. (2023). Genetic regulation of human isomiR biogenesis. *Cancers*, 15(17), 4411.
3. Schneider, B. P., **Jiang, G**., Ballinger, T. J., Shen, F., Chitambar, C., Nanda, R., ... & Radovich, M. (2022). BRE12-158: A Postneoadjuvant, Randomized Phase II Trial of Personalized Therapy Versus Treatment of Physician's Choice for Patients With

Residual Triple-Negative Breast Cancer. *Journal of Clinical Oncology*, 40(4), 345-355. (**co-first author**)

4. Radovich, M., **Jiang, G.**, Hancock, B. A., Chitambar, C., Nanda, R., Falkson, C., ... & Schneider, B. P. (2020). Association of circulating tumor DNA and circulating tumor cells after neoadjuvant chemotherapy with disease recurrence in patients with triple-negative breast cancer: preplanned secondary analysis of the BRE12-158 randomized clinical trial. *JAMA oncology*, *6*(9), 1410-1415. (**co-first author**)
5. Shen, F., **Jiang, G**., Philips, S., Gardner, L., Xue, G., Cantor, E., ... & Schneider, B. P. (2023). Cytochrome P450 Oxidoreductase (POR) Associated with Severe Paclitaxel-Induced Peripheral Neuropathy in Patients of European Ancestry from ECOG-ACRIN E5103. *Clinical Cancer Research*, *29*(13), 2494-2500. (**co-first author**)
6. Feng, S. Y., Liang, G. F., Xu, Z. S., Li, A. F., Du, J. X., Song, G. N., ... & **Jiang, G**. (2018). Meta-analysis of antiviral protection of white spot syndrome virus vaccine to the shrimp. *Fish & shellfish immunology*, 81, 260-265. (IF 3.185, **corresponding author**)
7. Bhat-Nakshatri, P., Gao, H., Khatpe, A. S., McGuire, P. C., Erdogan, C., Chen, D., **Jiang, G**., ... & Nakshatri, H. (2024). Single nuclei chromatin accessibility and transcriptomic map of breast tissues of women of diverse genetic ancestry., *Nature Medicine*, (*Accepted*)
8. Schneider, B. P., Zhao, F., Ballinger, T.J., Garcia, S.F., …, Wolff, A.C., (2024). ECOG-ACRIN EAZ171: Prospective validation trial of germline predictors of taxane-induced peripheral neuropathy (TIPN) in Black women with early-stage breast cancer., *Journal of Clinical Oncology*, (*Accepted*)
9. Shen, F., **Jiang, G**., Philips, S., Cantor, E., Gardner, L., Xue, G., ... & Schneider, B. P. (2024). Germline predictors for bevacizumab induced hypertensive crisis in ECOG-ACRIN 5103 and BEATRICE. *British journal of cancer*, 1-8.
10. Motolani, A., Martin, M., Wang, B., **Jiang, G**., Alipourgivi, F., Huang, X., ... & Lu, T. (2023). Critical Role of Novel O-GlcNAcylation of S550 and S551 on the p65 Subunit of NF-κB in Pancreatic Cancer. *Cancers*, *15*(19), 4742.
11. Noonan, M. L., Ni, P., Solis, E., Marambio, Y. G., Agoro, R., Chu, X., **Jiang, G**.,… & White, K. E. (2023). Osteocyte Egln1/Phd2 links oxygen sensing and biomineralization via FGF23. *Bone Research*, *11*(1), 7.
12. Prabhu, L., Martin, M., Chen, L., Demir, Ö., Jin, J., Huang, X., **Jiang, G**., ... & Lu, T. (2023). Inhibition of PRMT5 by market drugs as a novel cancer therapeutic avenue. *Genes & Diseases*, *10*(1), 267-283.
13. Wu, X., Shen, F., **Jiang, G**., Xue, G., Philips, S., Gardner, L., ... & Schneider, B. P. (2022). A non-coding GWAS variant impacts anthracycline-induced cardiotoxic phenotypes in human iPSC-derived cardiomyocytes. *Nature Communications*, *13*(1), 7171.
14. Cantor, E. L., Shen, F., **Jiang, G**., Tan, Z., Cunningham, G. M., Wu, X., ... & Schneider, B. P. (2022). Passage number affects differentiation of sensory neurons from human induced pluripotent stem cells. *Scientific Reports*, *12*(1), 15869.
15. Ballinger, T. J., **Jiang, G**., Shen, F., Miller, K. D., Sledge Jr, G. W., & Schneider, B. P. (2022). Impact of African ancestry on the relationship between body mass index and survival in an early-stage breast cancer trial (ECOG-ACRIN E5103). *Cancer*.

16. Martin, M., Mundade, R., Hartley, A. V., **Jiang, G**., Jin, J., Sun, S., ... & Lu, T. (2022). Using VBIM technique to discover ARMC4/ODAD2 as a novel negative regulator of NF-κB and a new tumor suppressor in colorectal cancer. *International Journal of Molecular Sciences*, *23*(5), 2732.
17. Wei, H., Hartley, A. V., Motolani, A., Jiang, G., Safa, A., Prabhu, L., ... & Lu, T. (2022). A complex signature network that controls the upregulation of PRMT5 in colorectal cancer. *Genes & Diseases*, *9*(2), 285.
18. Quintanilha, J. C., Wang, J., Sibley, A. B., Jiang, C., Etheridge, A. S., Shen, F., **Jiang, G**., ... & Innocenti, F. (2022). Bevacizumab-induced hypertension and proteinuria: a genome-wide study of more than 1000 patients. *British Journal of Cancer*, *126*(2), 265-274.
19. Cunningham, G. M., Shen, F., Wu, X., Cantor, E. L., Gardner, L., Philips, S., **Jiang, G**., ... & Schneider, B. P. (2022). The impact of SBF2 on taxane-induced peripheral neuropathy. *PLoS genetics*, *18*(1), e1009968.
20. Cardenas, H., Fang, F., Jiang, G., Perkins, S. M., Zhang, C., Emerson, R. E., ... & Nephew, K. (2021). Methylomic signatures of high grade serous ovarian cancer. *Epigenetics*, *16*(11), 1201-1216.
21. Narasimhan, A., Zhong, X., Au, E. P., Ceppa, E. P., Nakeeb, A., House, M. G., **Jiang, G**., ... & Zimmers, T. A. (2021). Profiling of Adipose and Skeletal Muscle in Human Pancreatic Cancer Cachexia Reveals Distinct Gene Profiles with Convergent Pathways. *Cancers*, *13*(8), 1975.
22. Ballinger, T. J., **Jiang, G**., Kassem, N., Radovich, M., & Schneider, B. P. (2021). Impact of Body Mass Index on Presence of ctDNA and Disease Recurrence after Neoadjuvant Chemotherapy for Triple-Negative Breast Cancer: Analysis from BRE12-158. *Clinical Cancer Research*, *27*(4), 1195-1199.
23. Hartley, A. V., Wang, B., Mundade, R., Jiang, G., Sun, M., Wei, H., ... & Lu, T. (2020). PRMT5-mediated methylation of YBX1 regulates NF-κB activity in colorectal cancer. *Scientific reports*, *10*(1), 1-14.
24. Liu, J., Dong, C., **Jiang, G**., Lu, X., Liu, Y., & Wu, H. (2020). Transcription factor expression as a predictor of colon cancer prognosis: a machine learning practice. *BMC medical genomics*, *13*, 1-10.
25. Kays, J. K., Koniaris, L. G., Cooper, C. A., Pili, R., **Jiang, G**., Liu, Y., & Zimmers, T. A. (2020). The combination of low skeletal muscle mass and high tumor interleukin-6 associates with decreased survival in clear cell renal cell carcinoma. *Cancers*, *12*(6), 1605.
26. Hartley, A. V., Wang, B., **Jiang, G**., Wei, H., Sun, M., Prabhu, L., ... & Lu, T. (2020). Regulation of a PRMT5/NF-κB Axis by Phosphorylation of PRMT5 at Serine 15 in Colorectal Cancer. *International journal of molecular sciences*, *21*(10), 3684.
27. Patel, J., Dustrude, E., Haulcomb, M., Li, L., **Jiang, G**., Liu, Y., ... & Shekhar, A. (2020). Role of PSD95 and nNOS Interaction in Gene Regulation following Fear Conditioning and Implications for Molecular Mechanisms Underlying Post-Traumatic Stress Disorder. *Biological Psychiatry*, *87*(9), S334.
28. Dong, C., Liu, J., Chen, S. X., Dong, T., Jiang, G., Wang, Y., ... & Liu, Y. (2020). Highly robust model of transcription regulator activity predicts breast cancer overall survival. *BMC Medical Genomics*, *13*, 1-10.

29. Cardenas, H., **Jiang, G**., Pepin, J. T., Parker, J. B., Condello, S., Nephew, K. P., ... & Matei, D. (2019). Interferon-γ signaling is associated with BRCA1 loss-of-function mutations in high grade serous ovarian cancer. *NPJ precision oncology*, *3*(1), 1-14. (co-first author)
30. Sahu, S. S., Dey, S., Nabinger, S. C., **Jiang, G**., Bates, A., Tanaka, H., ... & Kota, J. (2019). The role and therapeutic potential of miRNAs in colorectal liver metastasis. *Scientific reports*, *9*(1), 1-10.
31. Naramore, S. K., Bennett Jr, W. E., Jiang, G., Kugathasan, S., Denson, L. A., Hyams, J. S., & Steiner, S. J. (2019). Serologic, but not genetic, markers are associated with impaired anthropometrics at diagnosis of pediatric Crohn's disease. *Journal of Pediatric Gastroenterology and Nutrition*, *69*(5), e129-e134.
32. Zhong X, Pons M, Poirier C, Jiang Y, Liu J, Sandusky GE, Shahda S, Nakeeb A, Schmidt CM, House MG, Ceppa EP, Zyromski NJ, Liu Y, **Jiang G**, Couch ME, Koniaris LG, Zimmers TA. The systemic activin response to pancreatic cancer—implications for effective cancer cachexia therapy. *Journal of Cachexia, Sarcopenia and Muscle*, 2019, 10(5):1083-1101.
33. Li, Y., Liu, Y., Xu, H., **Jiang, G**., Van der Jeught, K., Fang, Y., ... & Luo, Z. (2018). Heterozygous deletion of chromosome 17p renders prostate cancer vulnerable to inhibition of RNA polymerase II. *Nature communications*, 9(1), 4394.
34. Wei, H., Prabhu, L., Hartley, A. V., Martin, M., Sun, E., Jiang, G., ... & Lu, T. (2018). Methylation of NF-κB and its Role in Gene Regulation. *In Gene Expression and Regulation in Mammalian Cells-Transcription From General Aspects.* IntechOpen.
35. Schneider BP, Shen F, **Jiang G**, O'Neill A, Radovich M, Li L, Gardner L, Lai D, Foroud T, Sparano JA, Sledge Jr GW. Impact of Genetic Ancestry on Outcomes in ECOG-ACRIN-5103. *JCO Precision Oncology*. 2017 Aug 21. (***co-first author***)
36. Ballinger, T. J., Kassem, N., Shen, F., Jiang, G., Smith, M. L., Railey, E., ... & Schneider, B. P. (2017). Discerning the clinical relevance of biomarkers in early stage breast cancer. *Breast cancer research and treatment*, *164*, 89-97.
37. Martin M, Hua L, Wang B, Wei H, Prabhu L, Hartley AV, **Jiang G**, Liu Y, Lu T. Novel Serine 176 phosphorylation of YBX1 activates NF-κB in colon cancer. *Journal of Biological Chemistry*. 2017 Feb 24;292(8):3433-44.
38. Fang, F., Cardenas, H., Huang, H., **Jiang, G**., Perkins, S.M., ZHANG, C., Keer, H.N., Liu, Y., Nephew, K.P. and Matei, D., 2017. Genomic and Epigenomic Signatures in Ovarian Cancer Associated with Re-sensitization to Platinum Drugs. *Cancer research*, pp. canres-1492.
39. **Jiang G**, Fu Y, Zhang P, Ardeshir-Rouhani-Fard S, Cheng L, Li L, Li Z. Expression quantitative locus mapping for identification of hotspots using an empirical Bayes mixture model. *International Journal of Computational Biology and Drug Design*. 2017;10(2):108-22.
40. Schneider BP, Shen F, Gardner L, Radovich M, Li L, Miller KD, **Jiang G**, Lai D, O'Neill A, Sparano JA, Davidson NE. Genome-wide association study for anthracycline-induced congestive heart failure. *Clinical Cancer Research*. 2017 Jan 1;23(1):43-51.
41. Schneider, B.P., Lai, D., She, F., **Jiang, G**., etc., Charcot-Marie-Tooth Gene, SBF2, Associated with Taxane-Induced Peripheral Neuropathy in African Americans., *Oncotarget*. 2016 Dec 13;7(50):82244.

42. **Jiang G**, Zhang S, Yazdanparast A, Li M, Pawar AV, Liu Y, Inavolu SM, Cheng L. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. *BMC genomics*. 2016 Aug 22;17(7):525.
43. Radovich M, Kiel PJ, Nance SM, Niland EE, Parsley ME, Ferguson ME, **Jiang G**, Ammakkanavar NR, Einhorn LH, Cheng L, Nassiri M. Clinical benefit of a precision medicine based approach for guiding treatment of refractory cancers. *Oncotarget*. 2016 Jul 15;7(35):56491-500.
44. Schneider, B.P., Li, L., Radovich, M., Shen, F., Miller, K.D., Flockhart, D.A., **Jiang, G**., Vance, G., Gardner, L., Vatta, M. and Bai, S., Genome-Wide Association Studies for Taxane-Induced Peripheral Neuropathy in ECOG-5103 and ECOG-1199. *Clinical Cancer Research*, 2015 Nov 15;21(22):5082-91.
45. Fournier, P.G., Juárez, P., **Jiang, G**., Clines, G.A., Niewolna, M., Kim, H.S., Walton, H.W., Peng, X.H., Liu, Y., Mohammad, K.S. and Wells, C.D., The TGF-β Signaling Regulator PMEPA1 Suppresses Prostate Cancer Metastases to Bone. *Cancer cell*. 2015 Jun 8;27(6):809-21.
46. Wang L, Athinarayanan S, **Jiang G**, Chalasani N, Zhang M, Liu W: Fatty acid desaturase 1 gene polymorphisms control human hepatic lipid composition. *Hepatology* 2015 Jan 1;61(1):119-28.
47. Schneider B, Li L, Shen F, Miller K, Radovich M, O'Neill A, Gray R, Lane D, Flockhart D, **Jiang G**: Genetic variant predicts bevacizumab-induced hypertension in ECOG-5103 and ECOG-2100. *British journal of cancer*, 2014 Sep 9;111(6):1241.
48. Chakraborty A, **Jiang G,** Wang Z, Boustani M, Liu Y, Skaar T, Li L. New aQTL SNPs for the CYP2D6 identified by a novel mediation analysis of genome-wide SNP arrays, gene expression arrays, and CYP2D6 activity. *BioMed research international*. 2013 Oct 22;2013.(co-first author)
49. Schneider BP, Gray R, Radovich M, Vance GH, Li L, **Jiang G**, Miller KD, Gralow J, Dickler MN, Cobleigh M: Prognostic and predictive value of tumor VEGF gene amplification in metastatic breast cancer treated with paclitaxel with and without bevacizumab; results from ECOG 2100 trial. *Clinical Cancer Research,* 2013 Mar 1;19(5):1281-9.