2021

# Analyzing the Role of Noncoding RNA-Derived RNAs in Human Cancer

Yulong Huang
*University of South Alabama*

2021

# Analyzing the Role of Noncoding RNA-Derived RNAs in Human Cancer

Yulong Huang
*University of South Alabama*

# Analyzing the Role of Noncoding RNA-Derived RNAs in Human Cancer

By

Yulong Huang

A thesis submitted in partial fulfillment of the requirements of the Honors College at University of South Alabama and the Bachelor of Sciences in the Biomedical Sciences Department

University of South Alabama

Mobile

May 2021

Approved by:

Digitally signed by Glen Borchert
DN: cn=Glen Borchert, o=USA
COM, ou=Dept of Pharmacology,
email=borchert@southalabama.edu, c=US
Date: 2021.04.15 14:55:09 -05'00'

Mentor: Dr. Glen Borchert

Committee Member: Dr. Raymond Langley

4/15/2021

Committee Member: Dr. David Bourrie

Kathy J. Cooke

Dean, Honors College

## DEDICATION

I would like to dedicate this thesis to my parents Li Zhou and Jingshan Huang, without whom I could not be pursuing medicine today. Thank you for raising me and guiding me through the craziness of the past twenty years of my life while constantly supporting all my endeavors. I love you both more than I can express with words alone.

## ACKNOWLEDGMENTS

First, I would like to thank my mentor Dr. Glen Borchert for supporting me in my academic endeavors for these last three years and counting. With his expertise in biology, he has given me the guidance that I greatly needed to complete this undergraduate research project.

I would like to thank Dr. Jingshan Huang for providing his extensive knowledge and expertise in the fields of bioinformatics and computer science.

I would like to further extend my gratitude to Dr. Mika Houserova and Mohan Kasukurthi for assisting me in data collection and analysis, script writing, and the writing and editing of my thesis itself.

I would like to thank Dr. Timothy Sherman for supporting this research project from day one.

I would like to thank Dr. Kathy Cooke and Dr. Douglas Marshall for their encouragement within the Honors College.

Finally, I would like to thank my committee members Dr. Ray Langley and Dr. David Bourrie for their mentorship, feedback, and other contributions to bettering my thesis project.

## ABSTRACT

To date, there are no known cures or effective preventative measures for cancer. Research has been conducted on small noncoding RNAs (ncRNAs) and noncoding RNA-derived RNAs (ndRNAs) as potential contributors to cancer suppression and/or proliferation. Tumorigenesis may be slowed by identifying and targeting some ndRNAs that are frequently presented in different cases of cancer, which may improve patient prognoses. To identify these ndRNAs, we utilized next-generation sequencing (NGS), a DNA/RNA sequencing technology, to sequence nucleotides in the human transcriptome. Then, we analyzed NGS data in accordance with information from The Cancer Genome Atlas (TCGA) and Sequence Read Archive (SRA) to identify ndRNAs of interest.

Our goal for this project is to identify ndRNAs that are highly expressed in five cancers (breast, lung, kidney, corpus uteri, brain) and ndRNAs that are highly expressed in controls. From here, we can determine which ndRNAs are differentially expressed. These ndRNAs may be significant biomarkers in earlier identification of cancer or the development of more curated therapies. For this project, we used Short Uncharacterized RNA Finder (SURFr) to process raw RNA sequencing (RNAseq) files obtained from TCGA. We then conducted a correlation analysis of TCGA data, followed by an analysis of SRA patient control data. Overwhelmingly, specific miRNA expressions were most elevated across all five cancers, notably miR-21. Whereas miRNAs that are differentially expressed across many cancers (miR-21 in this study) may be significant biomarkers in cancer research, differentially expressed miRNAs in certain demographics or cancers (miR-29a, miR-205, and miR-9 in this study) may be biomarkers for specific cancers.

## TABLE OF CONTENTS

## LIST OF ABBREVIATIONS

AJCC: American Joint Committee of Cancer

API: Application Programming Interface

ASC: Alabama Supercomputer

BAM: Binary Alignment Map

BLAST: Basic Local Alignment Search Tool

CTC: circulating tumor cell

DEV: differentially expressed vector

DNAseq: DNA sequencing

FASTA: fast-all

GBC: gallbladder cancer

HPC: High-Performance Computing

mRNA: messenger RNA

miRNA: microRNA

miRNAseq: microRNA sequencing

MN: myeloid neoplasm

NCBI: National Center for Biotechnology Information

ncRNA: noncoding RNA

ndRNA: noncoding RNA-derived RNA

NGS: next-generation sequencing

NoSQL: non-Structured Query Language

NSCLC: non-small cell lung cancer

nt: nucleotide

PC: personal computer

PCR: polymerase chain reaction

RNAseq: RNA sequencing

rRNA: ribosomal RNA

sdRNA: small nucleolar RNA-derived RNA

snRNA: small nuclear RNA

snoRA: H/ACA box snoRNA

snoRD: C/D box snoRNA

SNORic: snoRNA in cancer

snoRNA: small nucleolar RNA

SQL: Structured Query Language

SRA: Sequence Read Archive

SURFr: Short Uncharacterized RNA Finder

SV: similarity vector

TCGA: The Cancer Genome Atlas

TIC: tumor-initiating cell

TNBC: triple-negative breast cancer

tRF: transfer RNA-derived fragment

tRNA: transfer RNA

## LIST OF FIGURES

## LIST OF TABLES

## INTRODUCTION

Cancer has been a rampant global problem affecting numerous patients for thousands of years. Scientists have long been searching for potential cures for cancer, though no known cures currently exist. Given the increasing number of poor diagnoses and prognoses for cancer patients, there has, as a result, been a shift towards a rapidly growing demand for improved treatment options. Within the last couple of decades, researchers have turned their focus towards studying potential oncogenic diagnostic and prognostic biomarkers, particularly in conjunction with the TNM staging system, where T denotes the primary tumor, N denotes the number of afflicted lymph nodes, and M denotes the metastatic quality (degree of spread) of cancer [1,2]. The TNM system was established in 1958 by the American Joint Committee of Cancer (AJCC) to provide a standard means by which cancer is diagnosed. This classification of malignant tumors estimates survival rates, provides treatments, and communicates information accurately among medical professionals. These three factors are determined independently of each other before being integrated into one of four stages of cancer, stages I–IV [1].

Emerging evidence strongly suggests that specific RNAs are strong biomarkers for the presence of cancer, with noncoding RNAs (ncRNAs) being one of the largest indicators of cancer proliferation [3,4]. Next-generation sequencing (NGS) is a new DNA/RNA sequencing technology in the field of bioinformatics that can be used to analyze several subclasses of ncRNAs for their prevalence in various types of cancer [2,5]. While there is still much that remains unknown or uncertain about the pathogenesis of cancer, continually improving research in this field provides an optimistic outlook for the future of cancer patients. Our research focuses on the pertinent regulating roles of ncRNA-derived RNAs (ndRNAs) in cancer proliferation. We believe that, through these studies, we can work towards generating improved prognoses for cancer patients.

Only 1% of the human genome is responsible for protein coding [6]. Despite this fact, researchers have uncovered that over 80% of the human genome is active. In the past, the portion of the human genome containing ncRNA was regarded as “junk DNA,” but new information regarding oncogenesis and drug resistance, detailed below, has surfaced within the last 10 years [3,4,6]. Interestingly, ncRNAs constitute roughly 99% of the mammalian RNA, though the exact amount of functional ncRNA transcripts remains unknown as their numbers are increasing with each passing year [6]. Three discrete classes of ndRNAs are microRNAs (miRNAs), transfer RNA (tRNA)-derived fragments (tRFs), and small nucleolar RNA (snoRNA)-derived RNAs (sdRNAs) [6,7,8]. Though once believed to be “junk DNA” excisions of ncRNAs, these three subclasses are fully functioning fragments of ncRNAs [6]. Dysregulations in ncRNA composition, as a result, can contribute to tumorigenesis [6,8]. Our proposed research, again, is to analyze the prevalence of ndRNAs in cancer, with all literature being focused on various RNA molecules, including ncRNAs [3,4,6,7,8], ndRNAs [7], miRNAs [2,6,9], tRFs [6], snoRNAs [3,4,6,7,8,10], and sdRNAs [7,8].

Beginning with the first class, miRNAs are 22 nucleotide (nt)-long sequences of ncRNAs that are suspected contributors to messenger RNA (mRNA) degradation. Due to the “seed sequence” located on nucleotides 2–7, miRNAs can interfere with cell expression and cell signaling pathways [2,6]. Through this mechanism, miRNAs directly contribute to the four main stages (development, progression, metastasis, and drug resistance) of cancer proliferation. Also, samples of plasma and serum from early stages of cancer have been found to contain miRNA fragments. Furthermore, miRNAs are highly specific and can be artificially engineered to target certain distinct genes in a variety of bodily sites, with a special focus on gastric, prostate, and breast cancers [2,6]. The second class of ncRNAs is categorized as tRFs, which are under 40 nt in length and have the potential to inhibit gene expression and regulate apoptosis. These tRFs are overexpressed in sex hormone-linked cell lines and serve as biomarkers for hormone-dependent cancers, such as breast and prostate cancers [6]. Finally, snoRNAs are 60–300 nt small ncRNAs responsible for post-transcriptional modification of ribosomal RNA (rRNA) [3,6]. In the context of cancer progression, snoRNAs are believed to participate in gene silencing similarly to miRNA gene silencing, especially with regards to breast, prostate, and lung cancers [6].

Several snoRNAs have been identified as potential biomarkers in cancer research, including C/D box and H/ACA box snoRNAs—snoRDs and snoRAs, respectively—snoRNA-42, and snoRNA-93 [3,4,6,8]. Given the emerging evidence surrounding tumorigenesis, information regarding cancerous genetic information and biomarkers was cataloged into The Cancer Genome Atlas (TCGA), a public-funded project developed in 2005 to index the expansive human genomic profiles of more than 30 different cancers [9]. Because several ndRNAs in oncogenes have yet to be discovered, the ndRNAs must be analyzed in an efficient manner [7]. Traditional sequencing techniques are too slow to process a large amount of data simultaneously. Thus, NGS was developed in response to this problem [2,5]. While NGS is still not being fully utilized within a clinical setting, its application provides invaluable information about genetic mutations in the pathogenesis of cancers [5]. In particular, breast cancer, prostate cancer, myeloid neoplasms, gallbladder cancer (GBC), and lung cancer are five subclasses of cancer that are of interest to researchers [2,3,4,5,6,8,10]. The poor outlook associated with cancer patients can be attributed to rapid proliferation, absence of targeted treatments, and late-stage diagnoses, often resulting in relapse after chemotherapy or surgery [2,10]. The key to preventing or slowing cancer development lies in targeting specific biomarkers of tumorigenesis, and early detection of the disease is crucial for improving the chances of long-term survival of patients [2]. As more progress is being made in ndRNA research, researchers are gradually growing closer to finding cancer-causing ndRNAs that may be biomarkers for treatment and developing potential cures for cancer.

Currently, one of the most heavily researched biomarkers for cancer is snoRNAs, a class of intronic ncRNAs [3,4]. Three of the most important factors performing regulation of snoRNA expression are host genes, copy number variation, and DNA methylation [3]. Furthermore, snoRNAs have the potential to alter the composition of rRNAs and small nuclear RNAs (snRNAs) [8]. As mentioned previously, snoRDs participate in 2’-O-methylation of targets and are overexpressed in prostate cancer and non-small cell lung cancer, while snoRAs aid in target pseudouridylation—a form of RNA epigenetic modification—and show overexpression in prostate cancer. Ribonucleoproteins were also shown to have a strong positive correlation with snoRNAs in the “co-activation and synergy” of cancer [3].

Non-small cell lung cancer (NSCLC) is a subtype of lung cancer that is believed to be caused by cancer stem cells known as tumor-initiating cells (TICs) [9]. TICs are suspected agents in lung cancer regeneration, although current treatment options target the tumor cells as a whole rather than the residual TICs themselves. Subsequently, relapse after cancer treatments is a common occurrence [10]. NSCLC is the primary cause of cancer-related deaths worldwide, with an average survival rate of 9.5 months post-diagnosis [4]. Part of the reason for the poor outlook surrounding lung cancer is that the disease frequently goes unnoticed until it is in its late stages, minimizing the chances of successful treatment. Mannoor et al. studied whether snoRNAs affect TICs by analyzing snoRNA expression in lung ALDH1+ and ALDH1– cells across 28 NSCLC tissues [10]. For this study, they used a quantitative polymerase chain reaction (PCR) to characterize snoRNA expression in 82 varying NSCLC tissues and implemented both *in vitro* and *in vivo* assays to determine whether snoRNAs contribute to the stem cell-like quality of TICs. Of the 28 NSCLC tissues, 22 were shown to have a presence of ALDH1+ cells (a type of TIC), and of the 22 snoRNAs, 21 were overexpressed for ALDH1+ relative to ALDH1– cells. Additionally, overexpression of snoRNA-3 and snoRNA-42 was found to be correlated with poor NSCLC patient prognoses. Moreover, snoRNA-42 was more highly expressed in CD133+ cells compared with CD133– cells. Thus, ALDH1+ and CD133+ are important biomarkers for lung cancer and lung TICs. The knockdown of snoRNA-42 corresponds with a decrease in cell proliferation of *in vitro* TICs [10]. Quantitative PCR further revealed that snoRNA-3 and snoRNA-42 directly impact patient survival rates, with snoRNA-42 being overexpressed in lung cancer tissues and seldom expressed in healthy tissues [4,10]. This snoRNA is located in lq22, a common genomic amplicon in NSCLC. Researchers Mei et al. used quantitative PCR methods to test both snoRNA-42 and KIAA0907, its host gene, for amplification in 10 NSCLC cell lines and one BEAS-2B cell line as a control, finding that snoRNA-42 was much more highly expressed in NSCLC cell lines than in its host gene [4]. The results of this study strongly suggest that lq22 targets snoRNA-42, which becomes overexpressed with lq22 amplification. A transfection test was conducted to determine cell viability with snoRNA-42 knockdown, or a reduction in the amount of snoRNA-42 present, and showed that a decrease in snoRNA-42 is correlated with decreased NSCLC cell proliferation and decreased *in vivo* and *in vitro* tumorigenicity. One method by which snoRNA-42 knockdown regulates tumorigenicity is in the activation of p53, a common target of "genetic inactivation in human cancer," which triggers NSCLC cell apoptosis [4]. Similarly, an increase in snoRNA-42 corresponds to increased cancer cell proliferation and growth, demonstrating its high potential as an oncogene of lung cancer [4]. Genetic mutations in snoRNAs are a contributor to tumorigenesis, and the above results strongly suggest that snoRNA-42 expression in TICs is heavily linked to the invasive nature of lung cancers [10].

Another form of malignancy that has been researched heavily is breast cancer. In a study conducted by Gong et al., researchers performed snoRNA analysis of multiple samples within 31 different cancers categorized in TCGA, with tests showing that snoRD-46 displayed clinical significance as a potential oncogene in breast cancer [3]. Furthermore, the presence of sdRNA-93 was studied by Patterson et al. in two distinct breast cancer cell lines—metastatic MDA-MB-231 and primary MCF-7—as well as in a control and three cancer subtypes—triple-negative breast cancer (TNBC) tumors, Luminal A tumors, and Luminal B Her2+ tumors. The MDA-MB-231 cell line was experimentally determined to display more invasive characteristics than the MCF-7 cell line. All three tumor subtypes exhibited varying degrees of sdRNA-93 expression, while the normal control tissue exhibited no traces of sdRNA-93 expression. Additionally, sdRNA-93 is an important regulator of a sarcosine protein known as Pipox, the presence of which is indicative of specific breast cancer malignancies [8]. Results from these experiments indicate that snoRNA-93 and sdRNA-93, its small nucleolar RNA derivative, promote breast cancer tumorigenesis [8].

Continuing forth, cancers of the biliary tract are particularly life-threatening diseases, a trait that is attributed to their rapid proliferation, absence of treatments, and late-stage diagnoses. GBC is a form of biliary tract cancer whose origins can be linked to diseases such as cholecystitis, typhoid infection, and wall calcification; older age; obesity; smoking; and so on. Additionally, females are more at risk for GBC in comparison with men. One precursor linked to GBC expression is loss of heterozygosity, the inability of a gene to express itself due to deletion. Variations in the KRAS oncogene, which has been shown to activate cell-cell signaling, were also shown to be prevalent in cases of GBC [2]. Furthermore, changes in the p53 gene were correlated with GBC, similarly to its active role in NSCLC [2,4]. The p53 gene has been shown to increase the likelihood of a benign tumor becoming malignant. Aggregates of the p53 protein were detected in several gallbladder tumors, while it was not expressed in healthy gallbladder tissue. According to Montalvo-Jave et al., cycloxygenase-2 expression is suspected to be a component of an “inciting inflammatory process” as yet another indicator of GBC pathogenesis [2]. Though cancer antigens can provide useful information for GBC development, their presence can provide potentially contradicting information, and thus they cannot be utilized as definitive diagnostic biomarkers. On the other hand, a more reliable diagnostic biomarker uses circulating tumor cells (CTCs), components of tumors that enter the bloodstream. A higher concentration of CTCs was found in GBC patients, and tests indicated that CTCs may be used to differentiate between benign and malignant biliary tract cancers. Besides diagnostic biomarkers, one of the most widely implemented prognostic biomarkers is the expression of various miRNAs, which have the potential to either progress or suppress cancer growth. Other less reliable but useful prognostic biomarkers are serum tumor markers and tyrosine kinase receptors, both of which are correlated with worse GBC prognosis [2].

Moreover, myeloid neoplasms (MNs) are another form of malignancy that is caused by proliferating cancerous myeloid cells. Common MN diseases include acute myeloid leukemia, myelodysplastic syndromes, myelodysplastic syndromes, and myeloproliferative neoplasms. For the clinical trial application used in this study, a custom onco-hematology score was developed and assigned to categorize genetic variants based on their relative levels of pathogenicity. Both germline gene variants and structural changes were useful in classifying the genetic information of MNs. Using this systematic categorization, 39 genes in 121 occurrences of MN displayed a total of 278 pathogenic variants, and 84% of patients displayed at least one variant [5]. Due to the grim outlook associated with both MNs and GBC, NGS techniques can be employed to target potential biomarkers to gain a more thorough understanding of the pathogenesis associated with myeloid and biliary tract cancers [2,5]. Because NGS can recognize more genetic mutations than traditional sequencing techniques, the utilization of NGS allows for the development of more targeted treatments for MN patients and, as a direct result, increases the likelihood of improved, more specific diagnoses and prognoses [5]. For instance, Gong et al. created “snoRNA in cancer (SNORic),” a data portal summarizing the results and analyses of these extensive studies, in hopes of promoting widespread, public access to these data [3]. With the SNORic portal, users can obtain information regarding which of the 1,524 snoRNAs studied across more than 10,000 samples are of clinical significance. With more progress being made in snoRNA studies, researchers are growing closer to finding targets of cancer-causing snoRNAs, thereby developing potential cures for relevant cancers [3]. Furthermore, Kasukurthi et al. developed a computational approach known as “Short Uncharacterized RNA Finder” (SURFr) for processing ndRNA data as it relates to oncogenesis. The tool was designed to process raw NGS files to output a set of ndRNAs potentially present in instances of patient cancer. Similarity Vectors (SVs) and Differential Expression Vectors (DEVs) facilitate instances of sequence alignment and pattern recognition and, respectively, were used to help classify novel ndRNAs [7]. MoVaK, the new alignment methodology that is part of the SURFr software, was able to process RNA sequencing (RNAseq) files in a mere fraction of the time (~1/100) of Basic Local Alignment Search Tool (BLAST), a traditional sequencing algorithm widely adopted in the bioinformatics research field. In a comparison, MoVaK averaged four minutes and 35 seconds to process each file, while BLAST averaged seven hours and 26 minutes to perform the same task. These results indicated that the methodology presented in this study has achieved significant improvement over conventional NGS data analysis tools, presenting a much faster, more efficient way of categorizing and analyzing ncRNAs and their derivatives in several cases of cancer [7].

More broadly speaking, NGS data can also be analyzed in accordance with the data in TCGA, a catalog of over 30 different cancers, to detect recurring patterns in multiple cases of patient cancer. When TCGA was first developed, the project was split into two phases: Phase I, a 3-year pilot run, focused on the testing and development of brain, lung, and ovarian cancers, while studies in phase II focused on the remainder of the 30 cancer subtypes. Some of the means by which TCGA processes datasets include RNAseq, microRNA sequencing (miRNAseq), and DNA sequencing (DNAseq). RNAseq analyzes information within RNA strands with high precision, and miRNAseq is a subcategory of RNAseq that analyzes the role that miRNAs play in cancer gene expression and regulation. Similarly, DNAseq examines insertions, deletions, mutations, and other changes within nucleotide sequences. NGS provides additional genomic data with regards to these malignancies [9]. Given the marked developments in cancer research, scientists are aiming to determine the potential oncogenes that contribute to cancer cell proliferation and growth by utilizing NGS and TCGA alongside other complementary resources.

The full scope of the significance of ndRNA regulation and expression in cancer cell lines has yet to be understood completely, generating a need for further research [3,4]. Despite recent advancements, more studies on diagnostic and prognostic biomarkers must be conducted to gain a deeper understanding of their implications. Continued research of GBC biomarkers, for instance, is essential to identifying more specific biomarkers for cancers and improving patient prognoses [2]. Currently, the main problem that exists with biomarkers is that they require further research before being integrated into the clinical setting. The incorporation of markers could conversely throw off the universal viability of the TNM system by introducing too many factors into the staging process. Many clinical studies have failed to produce very consistent data due to having small patient sample sizes [1]. Nevertheless, biomarkers have the potential to be extremely useful in characterizing cancers and can be utilized to create more targeted treatments, especially in conjunction with NGS [1,2,5,6,7]. While NGS is a powerful sequencing tool, it cannot be independently utilized yet. However, its application undoubtedly provides a deeper understanding of genetic mutations leading to the growth and proliferation of cancer cells [5].

TCGA provides an extensive database of cancer genomic data, and as more information is discovered, more preventative measures can be taken towards cancer as a whole [9]. Its functionality is significantly more efficient than that of traditional sequencing methods [7]. Two main issues that currently exist in this research, however, lie in biological barriers and targets of drug delivery. Nevertheless, these emerging studies present great potential for specific targeting of cancerous genes, and the therapeutic use of ncRNAs and their derivatives is consistently being improved upon by scientists [6]. Ultimately, as methods for ncRNA and ndRNA processing continually grow to become more efficient, researchers hope to be able to isolate specific cures for different cancer subtypes in the near future [9].

## **EXPERIMENTAL METHODS***Research goals and methodology overview*

The presence of ncRNAs and ndRNAs—especially miRNAs, sdRNAs, and tRFs—have been previously shown to indicate cancer proliferation [2,3,4,6,8]. Thus, our three research goals are as follows: (1) to identify highly expressed ndRNAs present in five different cancers: breast, lung, kidney, corpus uteri, and brain; (2) to identify highly expressed ndRNAs present in control data; and (3) to compare highly expressed ndRNAs from cancer and control datasets to determine which ndRNAs are differentially expressed. We categorize RNAs as being differentially expressed if they appear only in either cancer or control data sets but not in both.

We used SURFr to process raw RNAseq files obtained from TCGA. Note that SURFr is a software developed at the University of South Alabama that is much more efficient than BLAST, a traditional sequencing data processing software. From the SURFr processing results, we then conducted an in-depth correlation analysis, followed by an analysis of relevant, publicly available SRA control patient data. Finally, we identified a set of ndRNAs of our interest and used them to aid in our research goals.

### *Details of methodology*

1. Data Sources

The primary source of data used for this study was taken from human miRNA-seq Binary Alignment Map (BAM)-format datasets from 53 different cancer subtypes (breast, lung, liver, brain, colon, prostate, stomach, tongue, heart, skin, etc.), including those with relatively few case numbers (<100 cases), in TCGA. Additional data for control patients (those without cancer) were obtained from the Sequence Read Archive (SRA). The data algorithmic input was obtained from the National Center for Biotechnology Information (NCBI).

2. Software and Hardware

Our project requires one High-Performance Computing (HPC) system (AKA supercomputer), one lab workstation, one computational algorithm known as SURFr, and one database known as MongoDB. Note that supercomputers are unique compared to traditional personal computers (PCs) and are specially designed to handle memory-intensive and computation-intensive tasks that cannot be efficiently handled by a traditional PC. All of the resources listed below, other than the HPC system, are provided to us by the computer science department at the University of South Alabama.

- Alabama Supercomputer (ASC): The ASC is a physical supercomputer located in Huntsville, AL, that allows users to access the required computational resources remotely upon request. It is provided to us by the state of Alabama.
- Lab workstation: The workstation (32 GB RAM) on which we performed data analysis is located in the University of South Alabama School of Computing Data Science Lab, room 3319. Data analysis was performed with the help of the SURFr algorithm, with SRA data being analyzed directly using SURFr.
- SURFr: Briefly, SURFr (http://salts.soc.southalabama.edu/surfr) is an extremely efficient computational algorithm with linear time and space complexities

designed to process raw RNAseq NGS files (in FASTA, FASTQ, or text format) directly into a list of ndRNAs, which serve as potential biomarkers for cancer.

- MongoDB: As opposed to a traditional Structured Query Language (SQL) database, MongoDB is a non-SQL (NoSQL) document database, meaning that it is more flexible, more user-friendly, and faster than an SQL database, especially when dealing with semi-structured data. This database was used to store, organize, retrieve, and compare all the results produced by our data analysis job using SURFr.

3. Research Design

1) Download all accessible miRNA-seq datasets from TCGA, where a total of 11,082 files are readily available in BAM format (https://portal.gdc.cancer.gov/). All miRNA-seq files were downloaded onto an external hard drive.
2) Transfer the downloaded files to ASC in multiple batches (~2,000 files per batch). The files were transferred to the ASC because converting BAM to FASTA format is time-consuming and computationally intensive, so this task was better suited for a supercomputer rather than a traditional PC.
3) Write a Bash script to convert BAM files to FASTA files using a C++ Application Programming Interface (API), BamTools (https://github.com/pezmaster31/bamtools).
4) Using SURFr, generate lists of ndRNAs from all the FASTA files (from step 3).
5) Using MongoDB, combine all ndRNA outputs (metadata and clinical data) into one master index file, a comparison of all ndRNAs across all files.

6) Perform a correlation analysis of ndRNAs (from Step 4) with the master index file (from Step 5). The master index file contains information such as cancer type, average DNA sequence, RNA type, number of RNAs, percent standard deviation, and average expression. The primary information we used from this file comes from the RNA type and average expression, which was utilized to determine the most highly expressed RNAs in each cancer.
7) Compare information from correlation analysis of TCGA cancer patient data (from step 6) to further analyze SRA control patient data.

## RESULTS & DISCUSSION

A relatively new, unexplored category is that of ndRNAs in their role as regulators of tumorigenesis. The three primary goals of this thesis project are as follows: to identify ndRNAs that are highly expressed in patients with breast, lung, kidney, corpus uteri, and brain cancer; to identify ndRNAs that are highly expressed in patients without cancer; and to determine which ndRNAs are differentially expressed.

One unanswered question that we seek to resolve is whether the expression of these ndRNAs potentially contributes either to cancer proliferation or suppression in different patients. A follow-up question is what the implications of certain ndRNAs within the same cancer and across various cancers are. We also seek to explore how these ndRNAs affect patients belonging to different demographics, including age, gender, ethnicity, race, and vital status. As such, another question we pose is whether different ndRNAs are more highly expressed or dysregulated in those of a certain demographic than those of a different one. If so, our next goal is to explore how this differential expression or dysregulation presents itself in the context of cancer development, that is, which ndRNAs influence cancer progression. By answering these questions, we seek to bring attention to a list of ndRNAs that can serve as potential biomarkers for further research in developing treatments for different cancers, especially within certain populations of people who may be more vulnerable to cancer, such as older adults.

Several studies have already correlated dysregulations in certain ndRNAs to the development of certain cancers [2,3,4,6,8,10]. In those studies, generally, the same ndRNAs either contribute to or fight against the proliferation of the same cancer subtype, while different ndRNAs analogously affect different cancer subtypes. Based on these studies, following up with our initial questions, we wanted to explore whether some of the same ndRNAs are highly expressed in many different patients, regardless of differences in age, gender, ethnicity, or race. Additionally, we wanted to discover whether, for different subcategories of cancer, some ndRNAs are differentially expressed for patients expressing different cancer subtypes.

While one ndRNA may enhance or inhibit oncogenesis in those belonging to one race, that same RNA may have no or opposite effects in those of another race or ethnicity [11,12,13,14,15,16,17]. There may also be higher instances of ndRNA dysregulations in many non-Caucasian races compared with Caucasians [11,12,13,14]. With relation to gender, some miRNAs and miRNA precursors have been shown to be more highly expressed in female cancer patients than male cancer patients, so it is believed that these ndRNAs may contribute to the increased incidence of certain cancers in some populations but not others [18,19,20]. Regarding age, as people become older, they are also more susceptible to increased somatic mutations that may begin as soon as middle age [21]. Such mutations can lead to senescence, or permanent cell arrest, which may potentially increase older populations' risk of developing cancer [21,22]. Thus, we aim to investigate if the same ndRNAs may have similar effects in people of the same demographic yet may have different effects in people of different demographics.

In many cases, the degree of polymorphisms in various ndRNAs or the expression of some ndRNAs seems to influence the pathogenicity of cancer and thus to the subsequent deadliness of cancer [14,22,23]. As such, these ndRNAs may affect the vital status of patients, that is, whether the cancer patient survives the disease.

To begin investigating these goals, the first question we needed to answer is how many ncRNAs are present in our dataset of TCGA patients. From there, we could verify which ndRNAs are also present. Next, we researched which ndRNAs are differentially expressed in healthy patients and differentially expressed in cancer patients. This way, we could begin to isolate which ndRNAs may or may not be contributing to cancer proliferation. For instance, if an ndRNA was found to be highly expressed in both the healthy patient and cancer patient datasets, then we deduced that that ndRNA may not contribute to cancer cell development or suppression. However, if it was highly expressed in cancer patients but not healthy patients, then we may deduce that that ndRNA has the potential to influence cancer cell growth. As demonstrated below, some of our findings have already been validated from existing experimental studies.

We first focused on the five cancer types with the highest prevalence of cases in TCGA: breast, 1208 files; bronchus and lung, 1091 files; kidney, 1035 files; corpus uteri, 572 files; and brain, 537 files. The total number of ncRNAs (and, thus, the total number of ndRNAs) in all five cancer types is as follows: breast, 6,667; bronchus and lung, 7,872; kidney, 8,330; corpus uteri, 8,481; brain, 6,206. From there, breast and lung cancers were subdivided further into ethnicity, race, age (divided according to decade), gender, and vital status. Kidney, corpus uteri, and brain cancer datasets were analyzed as a whole, without the subdivisions mentioned above. We filtered all the results by three types of RNAs: miRNAs, snoRNAs, and tRNAs. To determine the prevalence of each particular RNA, the data were sorted in the ascending order of percentage standard deviation values, where the lowest standard deviation represents that particular RNA being the most common across many patient files, and the highest standard deviation represents that particular RNA being the least common across many patient files. Percentage standard deviation alone is insufficient to determine the significance or prevalence of an ndRNA for multiple reasons, one of which is that standard deviation values can be low for one file simply due to consistently low expression. Another reason that percent standard deviation alone is an inadequate measure for ndRNA data analysis is that standard deviation values can be relatively higher, despite low expression in most or all patient files. For this reason, furthermore, the average expressions of each RNA across different files were also calculated to determine how highly expressed each RNA is. Within each cancer type, the top ten most highly expressed RNAs belonging only to the categories of miRNAs, snoRNAs, or tRNAs were analyzed. These RNAs were subsequently compared with existing literature to determine if they have been previously researched as contributing to the development or treatment of cancer.

Across all five cancer sites and their data subdivisions, tRNAs were by far the most commonly found derived RNA type, with only some exceptions. That more tRNAs are so common compared to other ndRNAs is unsurprising since tRNAs are directly involved in protein synthesis and would be expected to present in higher amounts [6]. Consistently, miRNAs were often the most highly expressed derived RNA type. Overall, miRNAs have been found to contribute greatly to cancer progression due to their ability to interfere with cell expression and cell signaling pathways [6]. Thus, given their roles as major diagnostic and prognostic biomarkers, that they are highly expressed in cancer patients is also unsurprising.

### *Breast cancer*

Because breast cancer is more prevalent in females than in males, we began with the TCGA dataset of females diagnosed with breast cancer. Considering the top ten most highly expressed RNAs, are all miRNAs (**Figure 1**). Starting with miR-21, which shows the highest average expression in female breast cancer patients, its upregulation has been correlated with the progression of breast cancer, including advanced stages and low rates of patient survival [24]. Other miRNAs that were highly expressed in our dataset that are suspected to play roles in breast cancer upregulation and/or metastatic growth are miR-182, miR-10b, and miR-99b [25,26,27]. miR-143, miR-22, miR-30a, miRNAs of the *let-7a* family (miR-let-7a1, miR-let-7a2, and miR-let-7a3), miR10a, and miR-148a have been found either to be downregulated in breast cancer cells or to regulate the continued growth of breast cancer cells in cancer patients [28,29,30,31,32,33,34]. High expression of these miRNAs indicates that they may play roles in either upregulating or downregulating breast cancer in breast cancer patients.

Using the same parameters and comparing these top ten most highly expressed RNAs from the female breast cancer dataset with the male breast cancer dataset, fascinatingly, the same ten ndRNAs are also highly expressed in male patients, albeit in a somewhat different order (**Figure 2**). The standard deviation values in both the male and female breast cancer datasets are also relatively similar to one another and are fairly low, indicating that not only are these RNAs highly expressed, but they are also consistently expressed across multiple different patients (**Table 1**). Combined, these data reveal that, regardless of gender differences, the same few ndRNAs may be presented as potential biomarkers for improved breast cancer treatments. Nevertheless, the slight differences in the ranking of ndRNA expression between females and males indicate that some miRNAs may be more effective targets for developing treatments in females than in males or vice versa.

Now, filtering the breast cancer dataset by ethnicity alone, the same ten ndRNAs in the female breast cancer dataset were found among both the Hispanic/Latino and non-Hispanic/non-Latino populations. Similar results were obtained from the ethnicity not reported category, except miR-375 instead of the *let-7* family of miRNAs was found among the top ten most highly expressed ndRNAs.

When filtering the breast cancer dataset by race alone, the top ten most highly expressed ndRNAs in the female breast cancer dataset were, for the most part, the same ten found in black/African American, white, American Indian/Alaska Native, and Asian breast cancer patients. In the black/African American breast cancer dataset, miR-10a was not among the top ten. In the American Indian/Alaska native breast cancer dataset, miR-30a, miR-10a, and the *let-7* family were omitted from the top ten. In the Asian breast cancer dataset, miR-30a and miR-10a were not found in the top ten, indicating that perhaps these RNAs are not the most ideal targets for cancer treatment in Asian breast cancer patients. However, even though these miRNAs were absent from the top ten, they were still overexpressed and quite close to the top ten. Again, similar results to the ethnicity not reported category were obtained from the race not reported category, with miR-375 being among the top ten most highly expressed ndRNAs rather than the *let-7* family of miRNAs.

Filtering the breast cancer dataset by age alone, the top ten ndRNAs from the female breast cancer dataset were found across all age groups (ages 20 and up), with the exception of the 90–99 cohort in which the *let-7* family of miRNAs was still highly expressed but not among the top ten.

Filtering the breast cancer dataset by vital status alone, both alive and dead patients overexpressed the same ten ndRNAs as the female breast cancer dataset, indicating that the presence of these ndRNAs is unlikely to act as a measure of whether one will survive the disease. The consistent expression of the same ndRNAs in breast cancer patients of varying populations indicates a strong possibility that these RNAs play potentially significant roles in the progression of breast cancer in general.

In addition to patients with cancer, normal human breast tissue controls (five files) taken from the SRA database were compared to TCGA data. Like TCGA patient files, the control data were sorted by descending value of expression. After filtering the results to show only miRNAs, snoRNAs, and tRNAs, the top ten most highly expressed RNAs were shown to be markedly different from those in the breast cancer patients (**Table 2**). In fact, interestingly, almost none of the top ten miRNAs found in the normal breast tissue files, with the exception of miR-148a, were found in the top ten miRNAs of breast cancer patients, suggesting that highly expressed miRNAs like miR-21 do play important regulating roles in cancer development.

Across all TCGA breast cancer datasets, regardless of ethnicity, race, age, gender, or vital status, the same group of ndRNAs is, for the most part, overexpressed in breast cancer patients and not expressed in normal healthy patients. From these findings, we may deduce that these ndRNAs are the ones that are most likely to contribute to breast cancer proliferation.

*Lung cancer*

The next cancer subtype of focus to this study is lung cancer. Lung cancer presents in many patients who smoke, and many people begin smoking during, or even before, early adulthood. As such, the dataset that we felt was most logical to begin data analysis is the lung cancer cohort of patients ages 30 through 39, the youngest available set of patients in TCGA who received a lung cancer diagnosis. Again, we sorted and filtered the data such that we would see results for the top ten most highly expressed RNAs for lung cancer patients in the 30–39 age range. However, since this cohort only has three files, some of the standard deviation values for the top ten values tended to be higher. Interestingly, seven of the ndRNAs that are highly expressed in the ages 30–39 lung cancer dataset are the same ones that are expressed in the female and male breast cancer datasets, namely miR-21, miR-22, miR-143, miR-148a, miR-182, miRNAs of the *let-7* family, and miR-99b; new ndRNAs include miR-203a, miR-29a, and miR-205 (**Figure 3**). Of these miRNAs, the ones that have been shown to be overexpressed in lung cancer patients are miR-21 and miR-205 [35,36,37]. The miRNAs that have been reported either to be under-expressed or to play regulatory roles in lung cancer development are miR-22, miR-143, miR-148a, miR-182, miRNAs of the *let-7* family, miR-99b, miR-203a, and miR-29a [38,39,40,41,42,43,44,45]. Overexpression of these eight miRNAs implies that they may be contributing to the regulation of lung cancer.

Comparing the age 40–49 lung cancer cohort to the 30–39 lung cancer cohort, they share seven ndRNAs in common, with miR-203, miR-29a, and miR-205 being omitted from the top ten values. Comparing the age 50–59 lung cancer cohort to the age 30–39 lung cancer cohort, they share eight ndRNAs in common, with miR-29a being omitted from the top ten values and miR-205 being entirely absent from the dataset. Comparing the age 60–69, 70–79, and 80–89 lung cancer cohorts to the age 30–39 lung cancer cohort, they share eight ndRNAs in common, this time with both miR-29a and miR-205 being entirely absent from these three datasets. The 90–99 lung cancer cohort was not considered as it only contained one patient file. With the gradual disappearance of miR-29a and miR-205 as one ages, these two miRNAs may be implicated in age-based lung cancer development. As miR-29a expression has been demonstrated to inhibit lung cancer development, its absence may contribute to the higher cases of lung cancer seen in older lung cancer patients [45]. Li et al. found in a sample of 32 NSCLC patients under age 60 and 40 NSCLC patients over age 60 that miR-29a demonstrated high expression in 14 patients of each group. In the NSCLC group of patients above age 60, miR-29a demonstrated low expression in 26 patients, whereas, in the NSCLC group of patients under age 60, miR-29a demonstrated low expression in 18 patients [46]. These results are consistent with our data indicating that miR-29a is not as highly expressed as one ages. On the other hand, high expression of miR-205 has been implied to enhance lung cancer proliferation, so its absence as one ages may lead to improved prognoses for older lung cancer patients [37]. Additionally, in an article by Zeng et al., miR-205 was expressed over twice as much in NSCLC patients 65 and under when compared with NSCLC patients over age 65 [47].

Filtering the lung cancer dataset by ethnicity alone, miR-29a and miR-205 were entirely missing from all datasets. Additionally, in the Hispanic/Latino lung cancer cohort, miR-203a was also missing, and miR-182 was not among the top ten most highly expressed ndRNAs as with the age 30–39 lung cancer group. In the non-Hispanic/non-

Latino lung cancer population, miR-203a was not within the top ten most highly expressed ndRNAs.

Filtering the lung cancer data by race alone, compared to the age 30–39 lung cancer group, once again, miR-29a and miR-205 were entirely missing from all datasets except for the Asian and race not reported categories. In the black/African American group, miR-203a was not among the top ten ndRNAs but was very highly expressed. All other top ten ndRNAs from the age 30–39 lung cancer group were among the top ten most highly expressed ndRNAs in all datasets categorized by race. This consistency suggests that these ndRNAs may be relevant biomarkers in the continued research of developing lung cancer treatments for patients across multiple demographics.

Filtering the lung cancer data by gender alone, miR-29a and miR-205 were entirely missing from both female and male datasets. Otherwise, compared to the age 30–39 lung cancer cohort, only miR-203a was missing from the female lung cancer file, indicating that miR-203a may not be a significant target for female lung cancer treatment.

Filtering the lung cancer data by vital status alone, miR-29a and miR-205 were entirely missing from both alive and dead patients' files. Otherwise, the other top ten highly expressed ndRNAs in the age 30–39 lung cancer patients were shared with both the alive and dead lung cancer patients. As such, while miR-29a and miR-205 may be biomarkers for indicating the presence of lung cancer, it does not seem that either miRNA alone contributes significantly to determining the vital statuses of lung cancer patients.

Additionally, normal human lung tissue controls (six files) taken from SRA were compared to TCGA data. Comparing the top ten most highly expressed ndRNAs from both SRA and TCGA datasets, just like the breast cancer and normal breast tissue comparison, only miR-148a is shared between control and cancer patient files (**Table 3**). However, miR-205, although not one of the most highly expressed ndRNAs among all lung cancer patients, was highly expressed in younger lung cancer patients. miR-205 is also among the most highly expressed ndRNAs in the control dataset. Regardless, based on existing literature, the data indicate the potential significance of miR-205 in contributing to lung cancer proliferation.

Looking at the raw TCGA data, miR-29a and miR-205 are entirely absent from almost all datasets but are highly prevalent in younger lung cancer patients. Based on TCGA and SRA information and existing literature, miR-29a and miR-205 may be implicated in playing significant roles in age-based lung cancer inhibition or proliferation, respectively. However, it is important to note that miR-205 was also highly expressed in normal lung tissue controls. Other ndRNAs that were highly expressed across almost all lung cancer cohorts are also thought to contribute to the development of lung cancer.

### *Kidney, corpus uteri, and brain cancer*

As opposed to breast and lung cancer, the remaining analyses on kidney, corpus uteri, and brain cancer will be less detailed than those of breast and lung cancer and will mostly be analyzed as datasets in their entirety rather than being grouped by different demographics. Of note, in kidney cancer patients, several of the top ten highly expressed ndRNAs were the same ones present in breast and lung cancer, including miR-21, miR-

22, miRNAs of the *let-7* family, and miR-143 (**Figure 4**). Similar trends were seen in corpus uteri cancer patients (**Figure 5**).

SRA control data for human normal kidney tissue (six files) were compared with TCGA kidney cancer data, and miR-148a and miR-101-1 were highly expressed in both healthy and cancer patients (**Table 4**). Similarly, when comparing SRA control data for human normal endometrial tissue (ten files) with TCGA corpus uteri cancer data, miR-148a was also found to be highly expressed in both healthy and cancer patients (**Table 5**).

In brain cancer patients, we see the largest deviation from this typical pattern of ndRNA expression seen in the other four cancer types discussed so far. Interestingly, miR-21 expression, while still one of the most highly expressed brain cancer ndRNAs, is significantly under-expressed compared to breast, lung, kidney, and corpus uteri patients (**Figure 6**). Instead, miRNAs of the miR-9 family are the most highly expressed, and overexpression of miR-9 has been suggested to contribute specifically to brain tumorigenesis [48,49]. This would account for the unusually high expression of the miR-9 family in brain cancer patients as a whole.

SRA control data for human normal brain tissue (six files) were compared with TCGA kidney cancer data. No miRNAs were shared between the two datasets, indicating that the miRNAs identified in brain cancer tissue may be relevant biomarkers for cancer detection (**Table 6**).

### *Final thoughts*

By far, the most highly expressed ndRNAs in TCGA datasets were consistently miRNAs, and the most highly expressed ndRNA in nearly every dataset is miR-21, with the exception of brain cancer. Given the data in combination with existing literature, it would seem that upregulation of miR-21 is strongly implicated in a wide variety of cancers, including but not limited to the ones listed in this thesis [24,35,36,50,51]. Of all the RNAs in our datasets, except for brain cancer, miR-148a was found to be highly expressed in all control datasets in addition to the cancerous datasets. However, this does not indicate that its expression is not significant to cancer progression as overexpression of miR-148a has been linked to regulation of breast and lung cancer progression [34,40]. Furthermore, of note, across all cancer subtypes, miRNAs were by far the most highly expressed RNA type, which further highlights the roles that miRNAs specifically play as either oncogenes or tumor suppressors [52].

Looking at control data, miR-29a was found in all five healthy patient datasets but absent from all five cancer patient datasets, with the exception of the younger lung cancer cohorts. Additionally, miR-24 and miR-378a were found in all healthy patient datasets except brain cancer and were absent from all five cancer patient datasets (**Table 7**).

## **CONCLUSION**

Cancer is one of the world’s deadliest diseases as it currently lacks effective cures and preventative measures. Recent research has demonstrated that ncRNAs and ndRNAs are potential contributors to cancer suppression and/or proliferation. Particularly, tumorigenesis may be slowed or stopped by identifying and targeting some ndRNAs. In addition, early detection of specific small ndRNAs may significantly improve cancer patients’ chances of long-term survival. Therefore, our research goals in this project are to identify ndRNAs that are highly expressed in breast, lung, kidney, corpus uteri, and brain cancers; to identify ndRNAs that are highly expressed in controls; and to identify ndRNAs that are differentially expressed in cancer or healthy patients. We used the SURFr algorithm to process raw RNAseq files from TCGA and then conducted a correlation analysis, after which we analyzed relevant SRA control patient data.

We found that miRNA expression was most elevated across all five cancers, and miR-21 was the most highly expressed miRNA in all cases except for brain cancer. Notably, the implication of miR-21 overexpression has been studied in various cancers and is thought to contribute to its increased progression in patients [24,35,36,50,51]. Another important observation from this research is that, for the most part, the same miRNAs were found in relatively similar amounts, save for miR-29a and miR-205 in lung cancer patients of different age groups and miR-9 in brain cancer patients. Existing literature demonstrates the potential role that miR-29a and miR-205 can play in age-related lung cancer proliferation or suppression [21,22,37,45,46,47]. Furthermore, miR-9 was notably most highly expressed in the brain cancer dataset, and its overexpression has been correlated with brain tumorigenesis [48,49].

Out of the TCGA cancer data and SRA control data, miR-148a was the primary ndRNA of note found to be shared between the two sets, hence its expression is non-differential. Its role in cancer progression is still being researched, but its overexpression may contribute to the suppression of breast and lung cancer cell growth [34,40].

As such, we concluded that, while most miRNAs from our data were presented in roughly the same amounts across patients of different demographics and different cancers, some miRNAs are specific to some demographic or cancer type. miRNAs that are highly expressed across many different cancers (miR-21 in this study) have the potential to be biomarkers for recognizing cancers and developing more generalized cancer therapies. miRNAs specific to demographic or cancer types (miR-29a, miR-205, and miR-9 in this study have the potential to be biomarkers for recognizing specific cancers and developing specific therapies.

Currently, the exact role that ndRNAs play in the progression of diseases such as cancer is still undetermined. As more studies relevant to the role of ndRNAs in cancer proliferation and suppression are conducted, more curated forms of cancer treatments and therapies may be developed. This area of cancer research is a relatively new, unexplored frontier that presents great potential for future work using similar or other novel data analysis methods. Additionally, as ndRNAs are believed to interfere with the function of coding RNAs, the role of ndRNAs in other non-cancerous diseases presents another area of future study. Overall, many mechanisms of ndRNAs remain yet unknown, and further study of ndRNAs presents significant potential to develop better targeted, more efficient treatments for both cancerous and non-cancerous diseases.

## REFERENCES

1. Ludwig, J. A., & Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer*, *5*(11), 845-856. https://doi.org/10.1038/nrc1739
2. Montalvo-Jave, E. E., Rahnemai-Azar, A. A., Papaconstantinou, D., Deloiza, M. E., Tsilimigras, D. I., Moris, D., Mendoza-Barrera, G. E., Weber, S. M., & Pawlik, T. M. (2019). Molecular pathways and potential biomarkers in gallbladder cancer: A comprehensive review. *Surg Oncol*, *31*, 83-89. https://doi.org/10.1016/j.suronc.2019.09.006
3. Gong, J., Li, Y., Liu, C. J., Xiang, Y., Li, C., Ye, Y., Zhang, Z., Hawke, D. H., Park, P. K., Diao, L., Putkey, J. A., Yang, L., Guo, A. Y., Lin, C., & Han, L. (2017). A Pan-cancer Analysis of the Expression and Clinical Relevance of Small Nucleolar RNAs in Human Cancer. *Cell Rep*, *21*(7), 1968-1981. https://doi.org/10.1016/j.celrep.2017.10.070
4. Mei, Y. P., Liao, J. P., Shen, J., Yu, L., Liu, B. L., Liu, L., Li, R. Y., Ji, L., Dorsey, S. G., Jiang, Z. R., Katz, R. L., Wang, J. Y., & Jiang, F. (2012). Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene*, *31*(22), 2794-2804. https://doi.org/10.1038/onc.2011.449
5. Carbonell, D., Suárez-González, J., Chicano, M., Andrés-Zayas, C., Triviño, J. C., Rodríguez-Macías, G., Bastos-Oreiro, M., Font, P., Ballesteros, M., Muñiz, P., Balsalobre, P., Kwon, M., Anguita, J., Díez-Martín, J. L., Buño, I., & Martínez-Laperche, C. (2019). Next-Generation Sequencing Improves Diagnosis, Prognosis

and Clinical Management of Myeloid Neoplasms. *Cancers (Basel), 11*(9). https://doi.org/10.3390/cancers11091364

6. Romano, G., Veneziano, D., Acunzo, M., & Croce, C. M. (2017). Small non-coding RNA and cancer. *Carcinogenesis*, *38*(5), 485-491. https://doi.org/10.1093/carcin/bgx026
7. Kasukurthi, M. V., Zhang, D., Houserova, M., Huang, Y., Tan, S., Ma, B., Li, D., Benton, R., Lin, J., Li, S., Borchert, G., & Huang, J. (2019). SURFr: Algorithm for Identification and analysis of ncRNA-derived RNAs. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1504-1507. https://doi.org/10.1109/BIBM47256.2019.8983074
8. Patterson, D. G., Roberts, J. T., King, V. M., Houserova, D., Barnhill, E. C., Crucello, A., Polska, C. J., Brantley, L. W., Kaufman, G. C., Nguyen, M., Santana, M. W., Schiller, I. A., Spicciani, J. S., Zapata, A. K., Miller, M. M., Sherman, T. D., Ma, R., Zhao, H., Arora, R., Coley, A. B., Zeidan, M. M., Tan, M., Xi, Y., & Borchert, G. M. (2017). Human snoRNA-93 is processed into a microRNA-like RNA that promotes breast cancer cell invasion. *NPJ Breast Cancer*, *3*, 25. https://doi.org/10.1038/s41523-017-0032-8
9. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn), 19*(1A), A68-77. https://doi.org/10.5114/wo.2014.47136
10. Mannoor, K., Shen, J., Liao, J., Liu, Z., & Jiang, F. (2014). Small nucleolar RNA signatures of lung tumor-initiating cells. *Mol Cancer*, *13*, 104. https://doi.org/10.1186/1476-4598-13-104

11. Telonis, A. G., & Rigoutsos, I. (2018). Race Disparities in the Contribution of miRNA Isoforms and tRNA-Derived Fragments to Triple-Negative Breast Cancer. *Cancer Res*, *78*(5), 1140-1154. https://doi.org/10.1158/0008-5472.CAN-17-1947

12. Li, E., Ji, P., Ouyang, N., Zhang, Y., Wang, X. Y., Rubin, D. C., Davidson, N. O., Bergamaschi, R., Shroyer, K. R., Burke, S., Zhu, W., & Williams, J. L. (2014). Differential expression of miRNAs in colon cancer between African and Caucasian Americans: implications for cancer racial health disparities. *Int J Oncol*, *45*(2), 587-594. https://doi.org/10.3892/ijo.2014.2469

13. Bhagirath, D., Yang, T. L., Tabatabai, Z. L., Shahryari, V., Majid, S., Dahiya, R., Tanaka, Y., & Saini, S. (2019). Role of a novel race-related tumor suppressor microRNA located in frequently deleted chromosomal locus 8p21 in prostate cancer progression. *Carcinogenesis*, *40*(5), 633-642. https://doi.org/10.1093/carcin/bgz058

14. Bhardwaj, A., Srivastava, S. K., Khan, M. A., Prajapati, V. K., Singh, S., Carter, J. E., & Singh, A. P. (2017). Racial disparities in prostate cancer: a molecular perspective. *Front Biosci (Landmark Ed)*, *22*, 772-782. https://doi.org/10.2741/4515

15. Gupta, I., Sareyeldin, R. M., Al-Hashimi, I., Al-Thawadi, H. A., Al Farsi, H., Vranic, S., & Al Moustafa, A. E. (2019). Triple Negative Breast Cancer Profile, from Gene to microRNA, in Relation to Ethnicity. *Cancers (Basel)*, *11*(3). https://doi.org/10.3390/cancers11030363

16. Chen, Q. H., Wang, Q. B., & Zhang, B. (2014). Ethnicity modifies the association between functional microRNA polymorphisms and breast cancer risk: a HuGE meta-analysis. *Tumour Biol*, *35*(1), 529-543. https://doi.org/10.1007/s13277-013-1074-7

17. Nassar, F. J., Talhouk, R., Zgheib, N. K., Tfayli, A., El Sabban, M., El Saghir, N. S., Boulos, F., Jabbour, M. N., Chalala, C., Boustany, R. M., Kadara, H., Zhang, Z., Zheng, Y., Joyce, B., Hou, L., Bazarbachi, A., Calin, G., & Nasr, R. (2017). microRNA Expression in Ethnic Specific Early Stage Breast Cancer: an Integration and Comparative Analysis. *Sci Rep*, *7*(1), 16829. https://doi.org/10.1038/s41598-017-16978-y

18. Pinto, R., Pilato, B., Ottini, L., Lambo, R., Simone, G., Paradiso, A., & Tommasi, S. (2013). Different methylation and microRNA expression pattern in male and female familial breast cancer. *J Cell Physiol*, *228*(6), 1264-1269. https://doi.org/10.1002/jcp.24281

19. Sharma, S., & Eghbali, M. (2014). Influence of sex differences on microRNA gene regulation in disease. *Biol Sex Differ*, *5*(1), 3. https://doi.org/10.1186/2042-6410-5-3

20. Guo, L., Zhang, Q., Ma, X., Wang, J., & Liang, T. (2017). miRNA and mRNA expression analysis reveals potential sex-biased miRNA expression. *Sci Rep*, *7*, 39812. https://doi.org/10.1038/srep39812

21. Sandiford, O. A., Moore, C. A., Du, J., Boulad, M., Gergues, M., Eltouky, H., & Rameshwar, P. (2018). Human Aging and Cancer: Role of miRNA in Tumor

Microenvironment. *Adv Exp Med Biol*, *1056*, 137-152. https://doi.org/10.1007/978-3-319-74470-4_9

22. Fang, C., Li, X. P., Gong, W. J., Wu, N. Y., Tang, J., Yin, J. Y., Li, X., Zhang, W., Zhou, H. H., & Liu, Z. Q. (2017). Age-related common miRNA polymorphism associated with severe toxicity in lung cancer patients treated with platinum-based chemotherapy. *Clin Exp Pharmacol Physiol*, *44 Suppl 1*, 21-29. https://doi.org/10.1111/1440-1681.12704
23. Mulrane, L., McGee, S. F., Gallagher, W. M., & O'Connor, D. P. (2013). miRNA dysregulation in breast cancer. *Cancer Res*, *73*(22), 6554-6562. https://doi.org/10.1158/0008-5472.CAN-13-1841
24. Yan, L. X., Huang, X. F., Shao, Q., Huang, M. Y., Deng, L., Wu, Q. L., Zeng, Y. X., & Shao, J. Y. (2008). MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *RNA*, *14*(11), 2348-2360. https://doi.org/10.1261/rna.1034808
25. Zhang, X., Ma, G., Liu, J., & Zhang, Y. (2017). MicroRNA-182 promotes proliferation and metastasis by targeting FOXF2 in triple-negative breast cancer. *Oncol Lett*, *14*(4), 4805-4811. https://doi.org/10.3892/ol.2017.6778
26. Ma, L. (2010). Role of miR-10b in breast cancer metastasis. *Breast Cancer Res*, *12*(5), 210. https://doi.org/10.1186/bcr2720
27. Zhao, Y. J., Song, X., Niu, L., Tang, Y., & Xie, L. (2019). Circulating Exosomal miR-150-5p and miR-99b-5p as Diagnostic Biomarkers for Colorectal Cancer. *Front Oncol*, *9*, 1129. https://doi.org/10.3389/fonc.2019.01129

28. Ng, E. K., Li, R., Shin, V. Y., Siu, J. M., Ma, E. S., & Kwong, A. (2014). MicroRNA-143 is downregulated in breast cancer and regulates DNA methyltransferases 3A in breast cancer cells. *Tumour Biol*, *35*(3), 2591-2598. https://doi.org/10.1007/s13277-013-1341-7

29. Zou, Q., Tang, Q., Pan, Y., Wang, X., Dong, X., Liang, Z., & Huang, D. (2017). MicroRNA-22 inhibits cell growth and metastasis in breast cancer via targeting of SIRT1. *Exp Ther Med*, *14*(2), 1009-1016. https://doi.org/10.3892/etm.2017.4590

30. Zhang, H. D., Jiang, L. H., Sun, D. W., Li, J., & Tang, J. H. (2017). miR-30a inhibits the biological function of breast cancer cells by targeting Notch1. *Int J Mol Med*, *40*(4), 1235-1242. https://doi.org/10.3892/ijmm.2017.3084

31. Mansoori, B., Mohammadi, A., Shirjang, S., Baghbani, E., & Baradaran, B. (2016). Micro RNA 34a and Let-7a Expression in Human Breast Cancers is Associated with Apoptotic Expression Genes. *Asian Pac J Cancer Prev*, *17*(4), 1887-1890. https://doi.org/10.7314/apjcp.2016.17.4.1887

32. Liu, C., Chen, Z., Fang, M., & Qiao, Y. (2019). MicroRNA let-7a inhibits proliferation of breast cancer cell by downregulating USP32 expression. *Transl Cancer Res*, *8*(5). https://doi.org/10.21037/tcr.2019.08.30

33. Ke, K., & Lou, T. (2017). MicroRNA-10a suppresses breast cancer progression via PI3K/Akt/mTOR pathway. *Oncol Lett*, *14*(5), 5994-6000. https://doi.org/10.3892/ol.2017.6930

34. Li, Q., Ren, P., Shi, P., Chen, Y., Xiang, F., Zhang, L., Wang, J., Lv, Q., & Xie, M. (2017). MicroRNA-148a promotes apoptosis and suppresses growth of breast cancer cells by targeting B-cell lymphoma 2. *Anticancer Drugs*, *28*(6), 588-595. https://doi.org/10.1097/CAD.0000000000000498

35. Bica-Pop, C., Cojocneanu-Petric, R., Magdo, L., Raduly, L., Gulei, D., & Berindan-Neagoe, I. (2018). Overview upon miR-21 in lung cancer: focus on NSCLC. *Cell Mol Life Sci*, *75*(19), 3539-3551. https://doi.org/10.1007/s00018-018-2877-x

36. Liu, Z. L., Wang, H., Liu, J., & Wang, Z. X. (2013). MicroRNA-21 (miR-21) expression promotes growth, metastasis, and chemo- or radioresistance in non-small cell lung cancer cells by targeting PTEN. *Mol Cell Biochem*, *372*(1-2), 35-45. https://doi.org/10.1007/s11010-012-1443-3

37. Li, J. H., Sun, S. S., Li, N., Lv, P., Xie, S. Y., & Wang, P. Y. (2017). MiR-205 as a promising biomarker in the diagnosis and prognosis of lung cancer. *Oncotarget*, *8*(54), 91938-91949. https://doi.org/10.18632/oncotarget.20262

38. Zhang, K., Li, X. Y., Wang, Z. M., Han, Z. F., & Zhao, Y. H. (2017). MiR-22 inhibits lung cancer cell EMT and invasion through targeting Snail. *Eur Rev Med Pharmacol Sci*, *21*(16), 3598-3604.

39. Xia, H., Sun, S., Wang, B., Wang, T., Liang, C., Li, G., Huang, C., Qi, D., & Chu, X. (2014). miR-143 inhibits NSCLC cell growth and metastasis by targeting Limk1. *Int J Mol Sci*, *15*(7), 11973-11983. https://doi.org/10.3390/ijms150711973

40. Li, Y., Deng, X., Zeng, X., & Peng, X. (2016). The Role of Mir-148a in Cancer. *J Cancer*, *7*(10), 1233-1241. https://doi.org/10.7150/jca.14616

41. Li, Y., Zhang, H., Zhao, C., Fan, Y., Liu, J., Li, X., Liu, H., & Chen, J. (2018). MiR-182 inhibits the epithelial to mesenchymal transition and metastasis of lung cancer cells by targeting the Met gene. *Mol Carcinog*, *57*(1), 125-136. https://doi.org/10.1002/mc.22741

42. Zhao, W., Hu, J. X., Hao, R. M., Zhang, Q., Guo, J. Q., Li, Y. J., Xie, N., Liu, L. Y., Wang, P. Y., Zhang, C., & Xie, S. Y. (2018). Induction of microRNA-let-7a inhibits lung adenocarcinoma cell growth by regulating cyclin D1. *Oncol Rep*, *40*(4), 1843-1854. https://doi.org/10.3892/or.2018.6593

43. Kang, J., Lee, S. Y., Kim, Y. J., Park, J. Y., Kwon, S. J., Na, M. J., Lee, E. J., Jeon, H. S., & Son, J. W. (2012). microRNA-99b acts as a tumor suppressor in non-small cell lung cancer by directly targeting fibroblast growth factor receptor 3. *Exp Ther Med*, *3*(1), 149-153. https://doi.org/10.3892/etm.2011.366

44. Jin, J., Deng, J., Wang, F., Xia, X., Qiu, T., Lu, W., Li, X., Zhang, H., Gu, X., Liu, Y., Cao, W., & Shao, W. (2013). The expression and function of microRNA-203 in lung cancer. *Tumour Biol*, *34*(1), 349-357. https://doi.org/10.1007/s13277-012-0556-3

45. Liu, X., Lv, X., Yang, Q., Jin, H., Zhou, W., & Fan, Q. (2018). MicroRNA-29a Functions as a Tumor Suppressor and Increases Cisplatin Sensitivity by Targeting

NRAS in Lung Cancer. *Technol Cancer Res Treat*, *17*, 1533033818758905. https://doi.org/10.1177/1533033818758905

46. Li, Y., Wang, Z., & Jing, R. (2017). MicroRNA-29a functions as a potential tumor suppressor through directly targeting CDC42 in non-small cell lung cancer. *Oncol Lett*, *13*(5), 3896-3904. https://doi.org/10.3892/ol.2017.5888

47. Zeng, Y., Zhu, J., Shen, D., Qin, H., Lei, Z., Li, W., Liu, Z., & Huang, J. A. (2017). MicroRNA-205 targets SMAD4 in non-small cell lung cancer and promotes lung cancer cell growth in vitro and in vivo. *Oncotarget*, *8*(19), 30817-30829. https://doi.org/10.18632/oncotarget.10339

48. Nass, D., Rosenwald, S., Meiri, E., Gilad, S., Tabibian-Keissar, H., Schlosberg, A., Kuker, H., Sion-Vardy, N., Tobar, A., Kharenko, O., Sitbon, E., Lithwick Yanai, G., Elyakim, E., Cholakh, H., Gibori, H., Spector, Y., Bentwich, Z., Barshack, I., & Rosenfeld, N. (2009). MiR-92b and miR-9/9* are specifically expressed in brain primary tumors and can be used to differentiate primary from metastatic brain tumors. *Brain Pathol*, *19*(3), 375-383. https://doi.org/10.1111/j.1750-3639.2008.00184.x

49. Chen, X., Yang, F., Zhang, T., Wang, W., Xi, W., Li, Y., Zhang, D., Huo, Y., Zhang, J., Yang, A., & Wang, T. (2019). MiR-9 promotes tumorigenesis and angiogenesis and is activated by MYC and OCT4 in human glioma. *J Exp Clin Cancer Res*, *38*(1), 99. https://doi.org/10.1186/s13046-019-1078-2

50. Feng, Y. H., & Tsao, C. J. (2016). Emerging role of microRNA-21 in cancer. *Biomed Rep*, *5*(4), 395-402. https://doi.org/10.3892/br.2016.747

51. Bautista-Sánchez, D., Arriaga-Canon, C., Pedroza-Torres, A., De La Rosa-Velázquez, I. A., González-Barrios, R., Contreras-Espinosa, L., Montiel-Manríquez, R., Castro-Hernández, C., Fragoso-Ontiveros, V., Álvarez-Gómez, R. M., & Herrera, L. A. (2020). The Promising Role of miR-21 as a Cancer Biomarker and Its Importance in RNA-Based Therapeutics. *Mol Ther Nucleic Acids*, *20*, 409-420. https://doi.org/10.1016/j.omtn.2020.03.003

52. Peng, Y., & Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*, *1*, 15004. https://doi.org/10.1038/sigtrans.2015.4

## APPENDICES

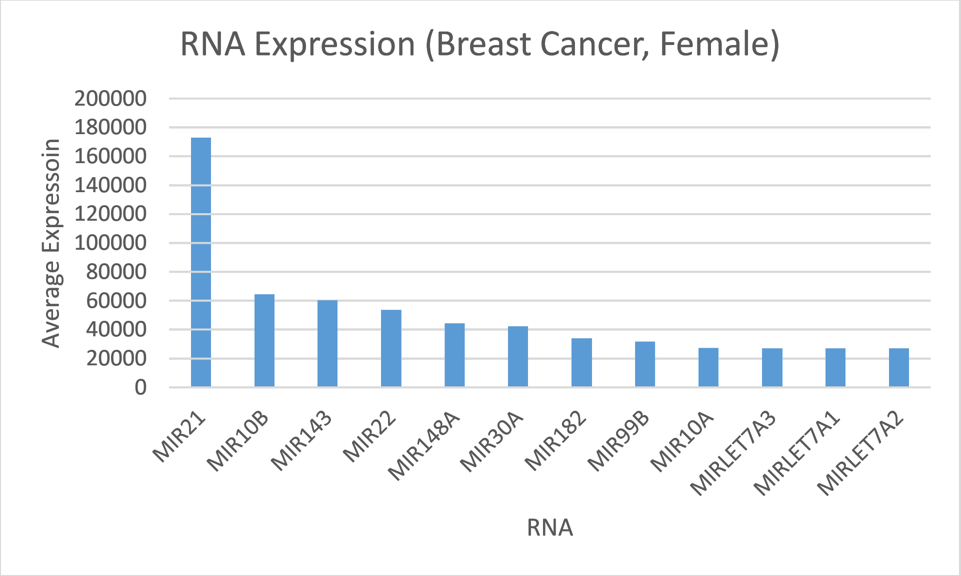

**Figure 1. Top ten most highly expressed ndRNAs in female breast cancer patients.** After filtering by miRNAs, snoRNAs, and tRNAs, the above results were obtained. miR-let7a1, miR-let7a2, and miR-let7a3 are taken to be the same RNA due to having identical average DNA sequences.

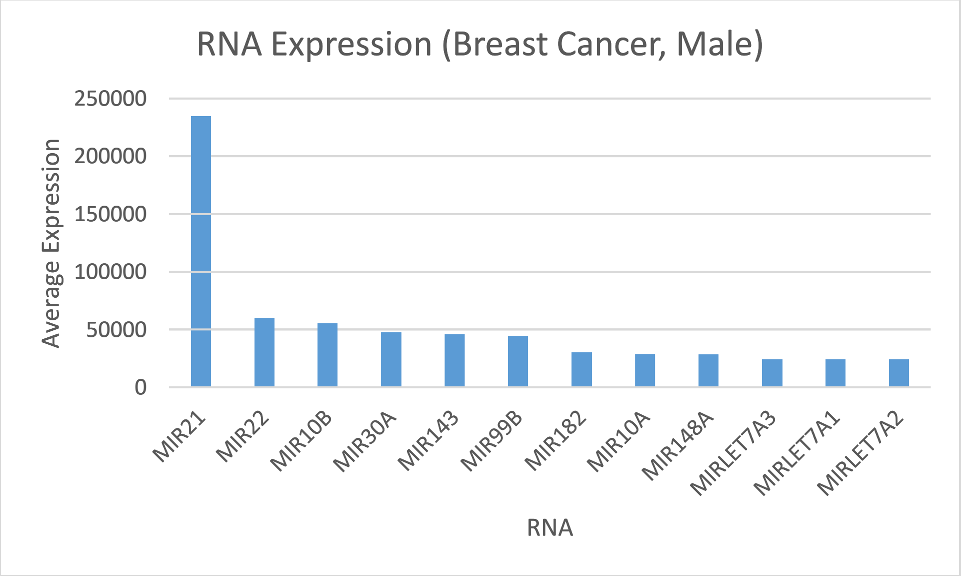

**Figure 2. Top ten most highly expressed ndRNAs in male breast cancer patients.** The same parameters applied to the female breast cancer patient dataset were applied to the male breast cancer patient dataset.

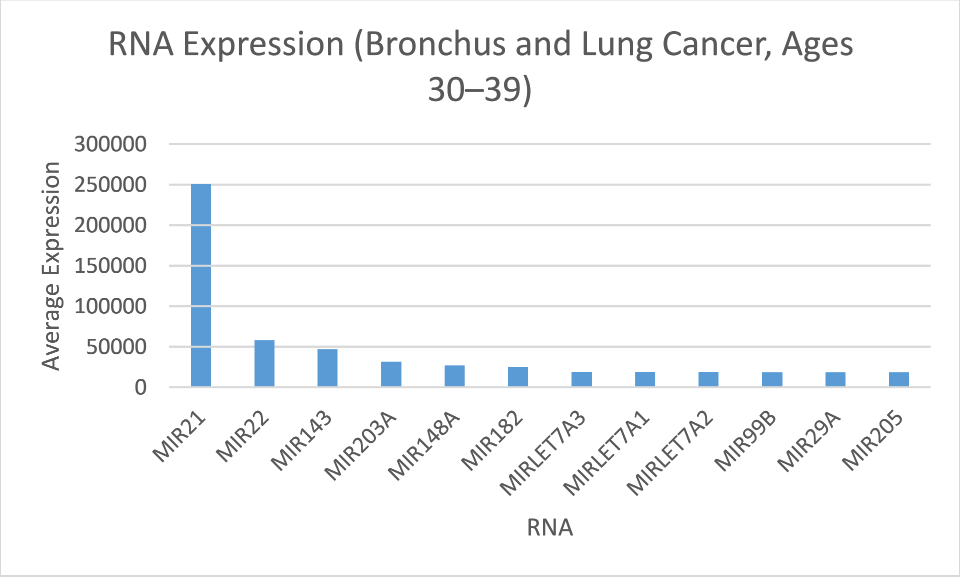

**Figure 3. Top ten most highly expressed ndRNAs in lung cancer patients, ages 30–39.** The same parameters applied to both breast cancer patient datasets were applied to the lung cancer patient datasets.

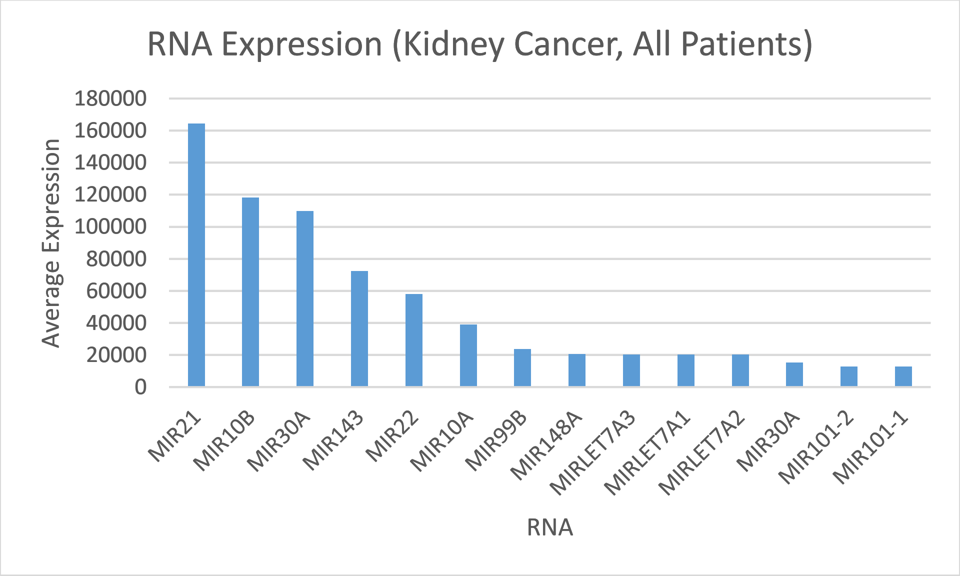

**Figure 4. Top ten most highly expressed ndRNAs in kidney cancer patients, all.** The same parameters applied to breast and lung cancer patient datasets were applied to kidney cancer patient datasets.

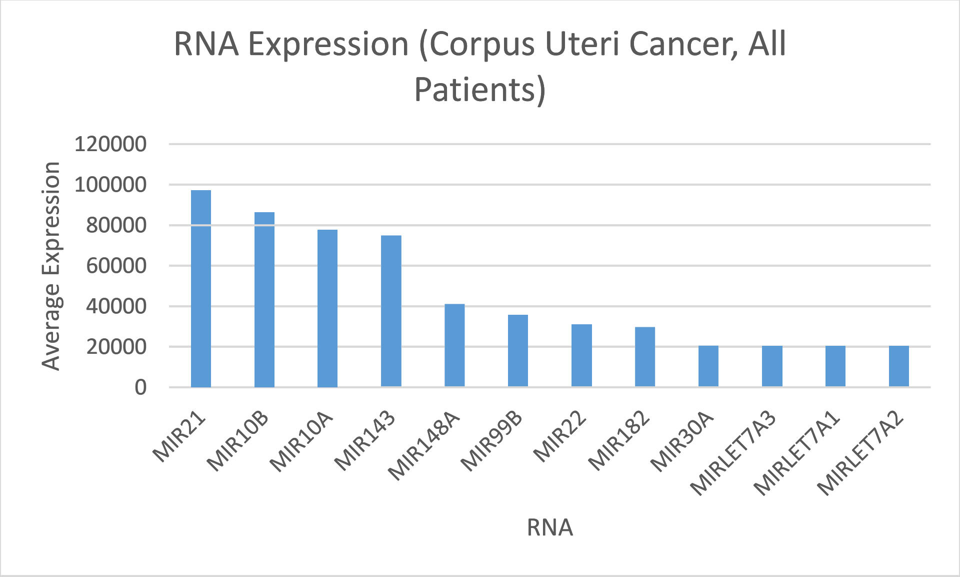

**Figure 5. Top ten most highly expressed ndRNAs in corpus uteri cancer patients, all.** The same parameters applied to breast, lung, and kidney cancer patient datasets were applied to corpus uteri cancer patient datasets.

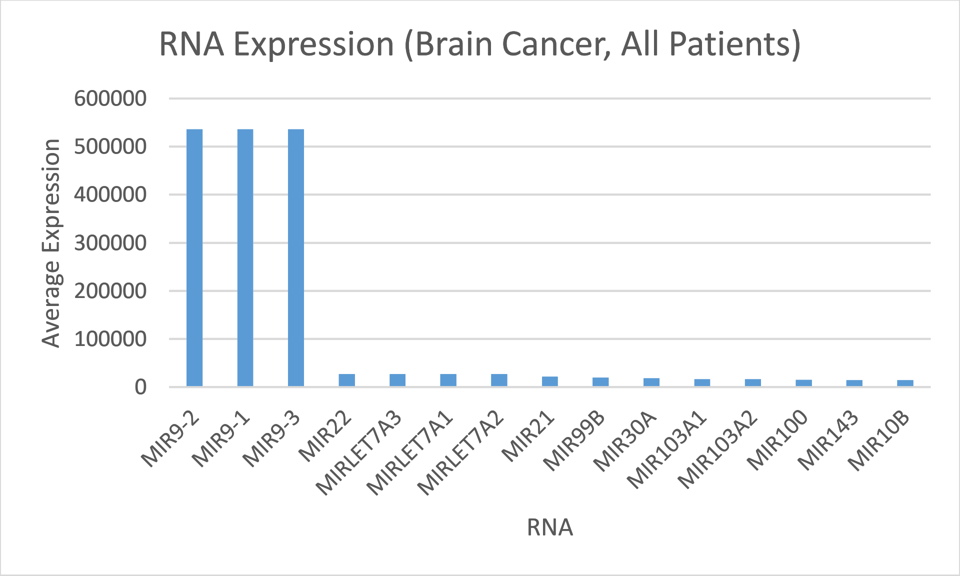

**Figure 6. Top ten most highly expressed ndRNAs in brain cancer patients, all.** The same parameters applied to breast, lung, kidney, and corpus uteri cancer patient datasets were applied to brain cancer patient datasets.

**Table 1. Comparison of standard deviation values of top ten most highly expressed ndRNAs between female and male breast cancer patients.** Of these most highly expressed ndRNAs in breast cancer patients, which are ranked in descending order of average expression, the percent standard deviations are listed below. Standard deviation values in both female and male breast cancer patients are relatively low. This indicates that these ndRNAs are not only highly expressed but are also *consistently* highly expressed across all TCGA breast cancer patient files.

***TCGA Breast Cancer Data***

| ***RNA*** | **Standard deviation, female** | **Standard deviation, male** |
|---|---|---|
| *MIR21* | 49.48 | 39.36 |
| *MIR10B* | 85.5 | 61.2 |
| *MIR143* | 84.36 | 61.1 |
| *MIR22* | 43.1 | 54.18 |
| *MIR148A* | 87.04 | 83 |
| *MIR30A* | 94.64 | 101.03 |
| *MIR182* | 77.13 | 59.69 |
| *MIR99B* | 69.43 | 57.44 |
| *MIR10A* | 134.87 | 126.48 |
| *MIRLET7A3* | 54.21 | 47.33 |
| *MIRLET7A1* | 54.27 | 47.43 |
| *MIRLET7A2* | 54.27 | 47.43 |

**Table 2. Comparison of top ten most highly expressed ndRNAs between breast cancer patients and breast tissue of normal patients.**

**Top Ten Highly Expressed RNAs**

| **Normal breast tissue** | **Breast cancer tissue** |
|---|---|
| MIR148A | MIR21 |
| MIR101-1 | MIR10B |
| MIR29A | MIR143 |
| MIR378A | MIR22 |
| SNORD66 | MIR148A |
| MIR24-2 | MIR30A |
| SNORD62 | MIR182 |
| SNORD3 | MIR99B |
| MIR222 | MIR10A |
| SNORD17 | MIRLET7A |

**Table 3. Comparison of top ten most highly expressed ndRNAs between lung cancer patients and lung tissue of normal patients.**

**Top Ten Highly Expressed RNAs**

| Normal lung tissue | Lung cancer tissue |
|---|---|
| MIR200C | MIR21 |
| MIR27A | MIR143 |
| MIR205 | MIR22 |
| MIR24-1 | MIR148A |
| MIR148A | MIRLET7A |
| MIR423 | MIR10A |
| MIR378A | MIR182 |
| MIR128-1 | MIR99B |
| MIR29A | MIR375 |
| MIR193A | MIR203A |

**Table 4. Comparison of top ten most highly expressed ndRNAs between kidney cancer patients and kidney tissue of normal patients.**

**Top Ten Highly Expressed RNAs**

| Normal kidney tissue | Kidney cancer tissue |
|---|---|
| MIR148A | MIR21 |
| MIR101-1 | MIR10B |
| MIR378A | MIR30A |
| MIR199A1 | MIR143 |
| MIR29A | MIR22 |
| MIR24-1,2 | MIR10A |
| MIR107 | MIR99B |
| SNORD48 | MIR148A |
| SNORD84 | MIRLET7A |
| MIR363 | MIR-101-1,2 |

**Table 5. Comparison of top ten most highly expressed ndRNAs between corpus uteri cancer patients and corpus uteri tissue of normal patients.**

| Top Ten Highly Expressed RNAs | |
|---|---|
| **Normal corpus uteri tissue** | **Corpus uteri cancer tissue** |
| MIR199A1 | MIR21 |
| MIR148A | MIR10B |
| tRNA-Gly-CCC-6-1 | MIR10A |
| MIR101-1 | MIR143 |
| tRNA-His-GTG-1 | MIR148A |
| MIR378A | MIR99B |
| SNORD52 | MIR22 |
| MIR29A | MIR182 |
| MIR24-1,2 | MIR30A |
| MIR107 | MIRLET7A |

**Table 6. Comparison of top ten most highly expressed ndRNAs between brain cancer patients and brain tissue of normal patients.**

| Top Ten Highly Expressed RNAs | |
|---|---|
| **Normal brain tissue** | **Brain cancer tissue** |
| MIR101-1 | MIR9 |
| MIR29A | MIR22 |
| SNORD115 | MIRLET7A |
| MIR29B2 | MIR21 |
| tRNA-Asp-GTC-2 | MIR99B |
| MIR107 | MIR30A |
| tRNA-Gln-CTG-5 | MIR103A |
| SNORD115-32 | MIR100 |
| SNORD2 | MIR143 |
| SNORD104 | MIR10B |

**Table 7. Prevalence of RNAs in all cancer and control data.** All miRNAs appearing in the top ten highly expressed RNAs of each cancer and control dataset were considered.

| | | *Cancer* | | | | | *Control* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Breast | Lung | Kidney | Corpus uteri | Brain | Breast | Lung | Kidney | Corpus uteri | Brain |
| ***miRNA*** | miR-9 | | | | | x | | | | | |
| | miR-10a | x | x | x | x | | | | | | |
| | miR-10b | x | | x | x | x | | | | | |
| | miR-21 | x | x | x | x | x | | | | | |
| | miR-22 | x | x | x | x | x | | | | | |
| | miR-24-1,2 | | | | | | x** | x* | x | x | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | miR-27a | | | | | | | x | | | |
| | miR-29a | | | | | | x | x | x | x | x |
| | miR-29b2 | | | | | | | | | | x |
| | miR-30a | x | | x | x | x | | | | | |
| | miR-99b | x | x | x | x | x | | | | | |
| | miR-100 | | | | | x | | | | | |
| | miR-101-1,2 | | | x | | | x | | x | x*** | x*** |
| | miR-103a | | | | | x | | | | | |
| | miR-107 | | | | | | | | x | x | x |
| | miR-128-1 | | | | | | | x | | | |
| | miR-143 | x | x | x | x | x | | | | | |
| | miR-148a | x | x | x | x | | x | x | x | x | |
| | miR-182 | x | x | | x | | | | | | |
| | miR-193a | | | | | | | x | | | |
| | miR-199-a1 | | | | | | | | x | x | |
| | miR-200c | | | | | | | x | | | |
| | miR-203a | | x | | | | | | | | |
| | miR-205 | | | | | | | x | | | |
| | miR-222 | | | | | | x | | | | |
| | miR-363 | | | | | | | | x | | |
| | miR-375 | | x | | | | | | | | |
| | miR-378a | | | | | | x | x | x | x | |
| | miR-423 | | | | | | | x | | | |
| | miR-let7a | x | x | x | x | x | | | | | |
| ***snoRNA*** | snoRD-2 | | | | | | | | | | x |
| | snoRD-3 | | | | | | x | | | | |
| | snoRD-17 | | | | | | x | | | | |
| | snoRD-48 | | | | | | | | x | | |
| | snoRD-52 | | | | | | | | | x | |
| | snoRD-62 | | | | | | x | | | | |
| | snoRD-66 | | | | | | x | | | | |
| | snoRD-84 | | | | | | | | x | | |
| | snoRD-104 | | | | | | | | | | x |
| | snoRD-115 | | | | | | | | | | x |
| | snoRD-115-32 | | | | | | | | | | x |
| ***tRNA*** | tRNA-Asp-GTC-2 | | | | | | | | | | x |
| | tRNA-Gln-CTG-5 | | | | | | | | | | x |
| | tRNA-Gly-CCC-6-1 | | | | | | | | | x | |
| | tRNA-His-GTG-1 | | | | | | | | | x | |

**only miR-24-1 present, **only miR-24-2 present, ***only miR-101-1 present*