UNIVERSITY OF THESSALY
FACULTY OF POSITIVE SCIENCES
DEPARTMENT OF COMPUTER SCIENCE AND BIOMEDICAL INFORMATICS

Application for conducting survival analysis utilizing the abundance of small RNAs

Maria-Anna Sotiropoulou

DIPLOMA THESIS
Supervisor
Prof. Artemis Hatzigeorgiou

Lamia, 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εφαρμογή για πραγματοποίηση ανάλυσης επιβίωσης αξιοποιώντας την αφθονία των μικρών RNAs

Μαρία-Άννα Σωτηροπούλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπουσα
Καθ. Άρτεμις Χατζηγεωργίου

Λαμία, 2021

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: …27…/..09…/2021……

Η Δηλούσα

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Εφαρμογή για πραγματοποίηση ανάλυσης επιβίωσης αξιοποιώντας την αφθονία των μικρών RNAs

Μαρία-Άννα Σωτηροπούλου

Τριμελής Επιτροπή:

Άρτεμις Χατζηγεωργίου, Καθηγήτρια (Επιβλέπουσα Καθηγήτρια)

Παντελεήμων Μπάγκος, Καθηγητής

Γεωργία Μπράλιου, Επίκουρος Καθηγήτρια

ABSTRACT

MicroRNAs (miRNAs) are small regulatory RNA molecules found in varying abundance in tissues and cell types. While their presence is indicative of the inhibitory regulation they exert on coding / non-coding transcripts, a number of studies have studied the diagnostic and prognostic dynamics of miRNAs. In the present work, the abundance of miRNAs was quantified by analyzing small RNA sequencing datasets of cancerous tissue from SRA and TCGA resources. Next, an application was designed to dichotomize a cohort of patients based on the abundance of small RNAs and perform comparative Kaplan-Meier survival analyses to elucidate prognostic biomarkers against clinical events (e.g., overall survival).

TABLE OF CONTENTS

1 Introduction

1.1 What are microRNAs?

In the last three decades, scientific research in the field of regulation of gene expression has experienced significant growth. The discovery of small RNA functional molecules which are not translated into proteins, contributed significantly to this. Groups of such "non-coding" RNAs are increasingly appearing to play a critical role in biological processes.

Non-coding RNAs found in eukaryotic cells, of which small RNAs constitute a family of regulatory RNAs, so far are classified into 4 subcategories depending on their function and form:

i. microRNAs [1]
ii. siRNAs (short interfering RNAs) [2]
iii. tncRNAs (tiny non-coding RNAs) [3]
iv. smRNAs (small modulatory RNAs) [4]

The topic of this thesis is focused on microRNAs. MicroRNAs (miRNAs) are short (~22 nucleotides (nt) long) endogenous non-coding RNA molecules. More specifically, in their mature state, miRNAs are 17-27 nt long and they have 5'-phosphate and 3'-hydroxyl ends [5-7] and are found in plants, animals and some viruses [8]. MiRNAs play essential roles in post-transcriptional regulation of gene expression and, thus, play a key role in development and diseases [9] as they are regulators of proliferation, differentiation, and cell death in both normal and aberrant pathways [10-14].

1.2 The first miRNA and their main function

The first identified miRNA was discovered in 1993 through a genetic screening in nematodes. It is the product a of *Caenorhabditis elegans* (*C. elegans*) gene and was called lin-4 [15]. The same month, the regulation of lin-14 by lin-4 was discovered [16]. Some years later, miRNA let-7 was discovered, also a product of *C. elegans* gene [17]. The main regulatory function of miRNAs was found to be the negative regulation of mRNA translation and its importance in controlling developmental timing was demonstrated [15, 18]. More specifically, when lin-4 or let-7 is inactivated, specific epithelial cells undergo additional cell divisions instead of their normal differentiation. Since abnormal cell proliferation is a hallmark of human cancers, it seems possible that miRNA expression patterns might denote the malignant state. Indeed, altered expression of numerous miRNAs has been found in some tumor types [19-22]. However, the first human disease associated with miRNA deregulation was chronic lymphocytic leukemia (CLL) [20]. In this disorder, miRNAs have a dual role working as both tumor suppressors and oncogenes [23]. Also, each miRNA is believed to regulate multiple genes, thus, more than one third of all human genes may be regulated by miRNA molecules [24].

1.3 Generation of miRNAs

miRNAs are produced from hairpin-shaped precursors, which are single stranded RNAs ~70 nt long. In animals, miRNAs genes are transcribed into long primary miRNAs (pri-miRNAs), which form hairpin-like structures. Processing of the pri-miRNAs is carried out by two RNase III type proteins, Drosha and Dicer [25]. Drosha processes the pri-miRNA hairpins in the nucleus

forming the precursor miRNAs (pre-miRNAs), which are transported by exportin-5 (EXP-5) in the cytoplasm. There, Dicer processes the pre-miRNAs in order to become mature miRNAs. After processing of the pre-miRNAs, mature miRNAs are tethered in Argonaute (Ago) subfamily proteins, to produce the effector RNA-induced silencing complex (RISC). RISC can target mRNAs and therefore functions as a post-transcriptional regulation system (Figure 1) [25].

In plants the two-step processing of primary miRNAs into mature ones is accomplished by a single RNase III enzyme, DCL1 (Dicer-like 1) [26].

1.4 Genome localization of miRNAs

Some miRNAs' genes are located in independent positions on the genome and some other are derived from introns of other genes. Also, some miRNAs' genes have their own promoters and some other share their protein-coding gene's promoter [3]. Moreover, many miRNAs' genes can be located at close distances from each other forming clusters, without necessarily having a shared functional relationship. These groups often have a common promoter and are being transcribed into a common primary transcript from which different miRNAs are derived.

1.5 miRNAs' Biosynthesis

Since the discovery of miRNAs, there has been a great effort to characterize their biosynthesis, as well as the mechanisms through which they are involved in biological processes that they regulate.

miRNAs are derived from endogenous transcripts whose transcription is achieved by both Polymerase II [27] and Polymerase III [28]. This depends on the location of the miRNAs' genes on the genome. For example, when their transcriptional regulation elements (the transcription factor binding sites and enhancers) are close to Alu repetitive sequences, the respective miRNA genes are transcribed by Polymerase III [29]. Furthermore, there are miRNAs whose genes are located inside introns or exons of protein coding genes. In this case miRNAs are not transcribed independently, but with their "host" gene and therefore they are regulated by the same elements [3, 30].

The Pol II/III-transcribed primary miRNA (pri-miRNA) forms a characteristic structure consisting of a cap at the 5' end, a polyadenylation tail and one or more hairpins (stem-loops) [31]. These primary molecules are thousands of nucleotides long, until with the help of a complex of proteins called microprocessor complex [32] they are cut in up to (into) 6 smaller sections of size approximately ~70-100nt, forming the precursors (pre-miRNAs) that have the structure of a hairpin. The microprocessor complex consists of Drosha, a type III RNase ribonuclease and a protein that binds double stranded RNA, called DGCR8 (Di George Syndrome Critical Region 8 Protein) (Figure 1) [33].

Pre-miRNAs are transferred from the nucleus to the cytoplasm by the protein Exportin 5 [34]. There, a second ribonuclease called Dicer cleaves the last base pairs and the loop. The result is a miRNA::miRNA* duplex, which consists of the mature miRNA and the complementary miRNA clone, "miRNA*". This duplex has a short lifespan and is degraded when the correct mature miRNA form is incorporated into the RNA-Induced Silencing Complex (RISC) [35, 36].

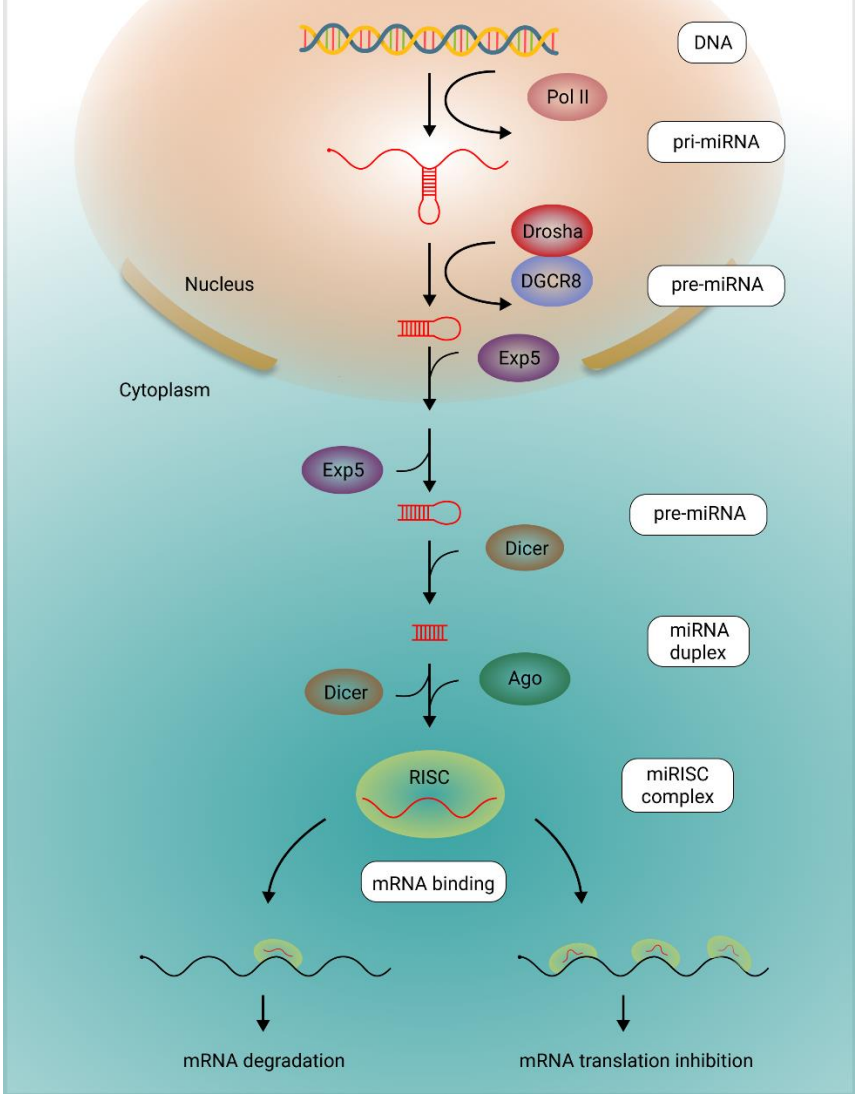

Figure 1: miRNA biosynthesis, image adapted from [37]

1.6 The cellular function of miRNAs

The post-transcriptional regulation of gene expression begins when the mature miRNA is attached to RISC, forming the miRISC complex. The central protein component of miRISC is a protein from the Argonaute (Ago) family, which catalyzes the miRNA-guided degradation of the mRNA itself, or the inhibition of its translation [38]. The mode of action of each miRNA depends not only on which member of the Argonaute is contained in the silencing complex, but more importantly the on base complementarity between the miRNA and the 3'UTR region it targets.

Initial studies showed that an almost perfect complementarity is required between miRNA and mRNA for degradation, and if this condition is met, then the miRISC complex deconstructs the target mRNA. Although this mode of function has been shown to occur in many organisms, it is mainly observed in plants, and very seldom in animals [35, 39].

Another function of mRNAs is the inhibition of translation, which occurs more often in animals. Although this function has been established, there are still many hypotheses about miRNAs' role in this mechanism. One of the most studied premises proposes that miRNAs have to interact with

proteins responsible for initiating translation into ribosomes, and more specifically, suppress the eukaryotic translation initiation factor 4E (eIF4E) [40]. This theory is supported by the fact that polysomes, in which mRNA has been targeted by miRNAs, are in an inactive state [41, 42]. Other theories state that repression occurs at a later stage from the start of translation [43], which is reinforced by data showing that polysomes are active during targeting by miRNAs [44]. These contradictory results possibly indicate that there are more than one translation inhibition mechanisms that depend on factors we are yet unaware of.

Limited data indicate that miRNAs may also have different functions besides mRNA degradation and translational inhibition. In some cases, it is possible that miRNAs bind to promoters of specific genes, inducing their expression, a phenomenon that has been named RNA activation (RNAa) [45].

1.7 Differentially abundant miRNAs in disease - potential biomarkers

miRNA regulation has been shown to affect many biological functions. In normal tissues, they are involved in functions such as embryogenesis, differentiation, programmed cell death and growth control. In addition, they seem to be a necessary element in normal cellular homeostasis, so if their concentration is disturbed, they can become responsible for the occurrence of diseases. As research proceeds, more and more miRNAs appear to be expressed both in different cell types and involved in specific dysfunctions.

For example, miRNA involvement in the regulation of differentiation, proliferation and apoptosis is now undisputed. In 2002, miRNAs were first associated with cancer and specifically with chronic lymphocytic leukemia (CLL). Two miRNA genes were found to be located in an area that had been deleted in 68% of the utilized samples [20]. From this study onwards, many miRNAs were found to possibly carry a role in oncogenesis or tumor suppression. The construction of miRNA signatures is a topic of intensive research, since specific miRNAs are expressed in each cancerous tissue. This fact can potentially facilitate grouping and diagnosis of many cancers [46].
Although most research is focused on the implication of miRNAs in cancer, their impact has also been associated with other pathologies. Studies have linked miRNAs to a variety of diseases such as heart disease [47], diseases of the nervous system, for example Schizophrenia and Alzheimer's disease [48], various types of diabetes [49], as well as other metabolic diseases [50]. It has even been accepted that miRNAs participate in immune defense mechanisms against viral and other infections [51].

1.8 Exploring the expression of miRNAs

miRBase is a sequence database that contains all published mature miRNA sequences, together with their predicted source hairpin precursors and annotation relating to their discovery, structure and function [52]. Since June 2003, when miRBase started cataloguing miRNA sequences, analysis of miRNA expression levels between different tissues, developmental stages, or disease states has been constant. Thus, miRNA expression levels were studied by several methods: Northern blots, real-time PCR, microarray analysis, in situ and solution hybridization. The most accurate and sensitive method from the aforementioned techniques is the quantitative reverse transcription PCR (qRT-PCR).

The main method used today to massively analyze the expression of miRNAs in tissues and different cell types is small RNA Sequencing (sRNA-Seq), which is based on the use of Next-

Generation Sequencing (NGS) technologies, enabling the simultaneous study of all captured miRNAs in a sample. There are many bioinformatics tools for miRNA quantification and the most commonly used one is miRDeep2. Specifically, miRDeep2 is a software package for identification of novel and known miRNAs in deep sequencing data, which can also be used for miRNA expression profiling across samples [53]. For analyzing miRNA NGS data, alignment of the short reads to the genome or the transcriptome is important. Currently, tools like Bowtie [54] are able to perform alignment efficiently and short-read aligners are always wrapped in miRNA quantification tools. miRDeep2 uses Bowtie to (i) map reads to the genome, (ii) map known miRNA precursors to the genome and (iii) to map known mature miRNAs onto their respective precursor sequences. This process allows it to derive the read counts of each known mature miRNA in a sample.

1.9 miRNAs as potential prognostic biomarkers – time-to-event analysis

Biological indicators (biomarkers) are medical signs or experimental measures of a biological state that are commonly used in populations, or in individuals, in order to monitor and predict health states, or to choose the most effective treatment course. More specifically, diagnostic biomarkers are used to indicate whether the subject is healthy or has a pathological disease, whereas predictive biomarkers are used to indicate the effectiveness of the specific treatment. Prognostic biomarkers can indicate the higher or lower possibility of a future clinical event in the group of individuals that is being studied.

A method of estimating time-to-event or survival models in order to examine the distribution of times between two events is Kaplan Meier (KM) estimate. 'Time-to-event' means the time from entry into a study until a subject has a particular outcome. The KM estimate is the simplest way of computing the survival over time in spite of all these difficulties associated with subjects or situations. [55] The "survival" term refers to the time from the beginning of a measurement until an event of interest happens, for example death. If a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up, the KM estimate takes these data into account and these are called censored data.

KM estimate is non-parametric and is used in survival analysis to analyze the data and to make comparisons between groups of participants, in our case splitting the patient cohort in two, using the log-rank test for hypothesis testing. The value of using the log-rank test is to assess whether the difference in survival between the two groups is statistically significant.

KM estimates survival probability as a factor of time, based on the occurrences of the event of interest at each point when any events occur. More specifically, the survival probability (S) at a specific time point (tx) corresponds to:

$$S_{t_x} = \frac{A_{t_0} - D_{t_x}}{A_{t_0} - C_{t_x}}$$

Where:
A: Alive subjects
D: Deceased subjects
C: Censored subjects

The cumulative (or total) probability of survival until any time point corresponds to the product of all survival probabilities before that time. Importantly, the time point at which the cumulative survival probability is 50% is the median survival time.

The KM curve is an estimator used to estimate the survival function. The KM curve is the visual representation of a function that shows the probability of an event, for example survival, at a respective time interval. In Figure 2, there are two curves, one representing the samples with high miRNA quantities and one the samples with low miRNA quantities. The original cohort was split in two in order to explore if patients' survival differs over time among the groups.

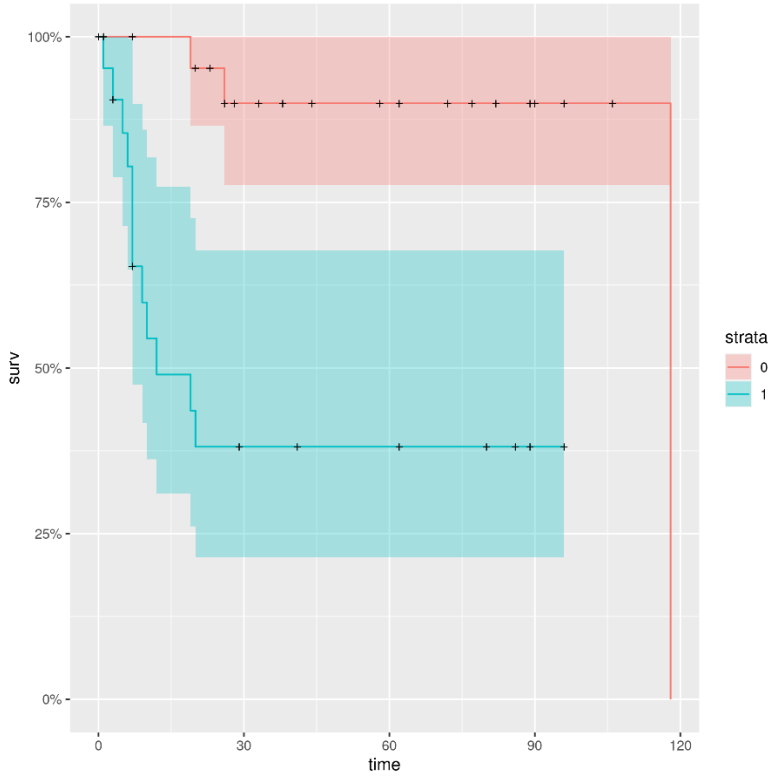

Figure 2: A Kaplan-Meier curve example, Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

miRNAs appear to be in varying amounts in different health states/conditions, hence it is important to study them as potential indicators. More specifically, miRNAs are being found in different amounts in healthy and pathological conditions or in a pathological condition in the course of a disease, so numerous studies focus on assessing their usefulness as prognostic markers that could be associated with an event.

Like all measurable indicators, miRNAs give us the opportunity to see if and which of them have the potential to act as prognostic biomarkers. Notably, the tissue-specific abundance of numerous miRNAs, their availability in body fluids, their stability against RNase digestion, and the various available methods enabling sensitive and inexpensive miRNA detection further support their study as biomarkers [56-59]. In light of the above, the involvement of miRNAs in several human diseases makes them potential diagnostic biomarkers, especially in cancer studies [57, 60]. Thus, numerous miRNA disease research tools have appeared [61, 62] and methods or database records to predict disease's diagnostic and prognostic miRNA biomarkers are being developed [63-65].

In this thesis's context, a tool for the assessment of prognostic capacity of miRNAs in sRNA-Seq data has been developed. It employs expression-based dichotomization of a given cohort and KM analysis for each miRNA. The prognostic assessment tool was experimentally tested using 2 cancer datasets, assessing patient death (overall survival) as an event.

2 Materials and Methods

2.1 Utilized Datasets

Two sets were analyzed and utilized in this thesis. The first set was generated derived from chondrosarcoma tissues of 102 patients, while the second referred to Estrogen Receptor positive (ER+) or negative (ER-) female breast cancer patients.

2.1.1 Chondrosarcoma Set

These datasets are deposited in the Sequence Read Archive (SRA), which is the largest publicly available repository of high throughput sequencing data and are available through multiple cloud providers and NCBI servers. SRA Runs are simply a manifest of data files that are linked to a given sequencing library. Each SRA Experiment is a unique sequencing library (SRA archive file) for a specific sample. The reason that these SRA Runs were chosen as an example of small RNA-Seq data before using the TCGA data is that SRA Runs were publicly available, had survival information and had less restrictions than the TCGA. The selected data are from the study "Integrated molecular characterization of chondrosarcoma reveals critical determinants of disease progression" of Rémy Nicolle et al. (ERP111275) [66] and are sRNA-Seq data from human chondrosarcoma tissues, which is a type of bone cancer. Sequencing was performed on Illumina HiSeq 2000 high-throughput sequencing platforms. The relevant supplementary data were also retrieved to obtain survival information for each patient. The survival analysis event was "overall survival" (patients' death) and the maximum follow-up time of patients (=max(time) months/12) was 14.91 years (max months 179).

2.1.2 Breast Cancer Set

The datasets for the Breast Cancer are sRNA-Seq data and their analysis was held by the DIANA-mAP tool [67] that automates the miRNA expression analysis steps that we did manually for the Chondrosarcoma dataset. These datasets' samples are from patients with breast cancer (TCGA-BRCA) and a filtering was performed so that patients:

1. were only female
2. did not have any other cancer at the same time (no other malignancies)

The two TCGA-BRCA datasets given for this application's tests were manually divided, depending on whether the tumors were classified as ER+ or ER-. This separation is related to whether the tumor expresses the Estrogen Receptor, so it is receptive to a number of treatments and generally has a better prognosis, or not. The ER- samples were 223 and the ER+ 725. The longest follow-up patients' time was 19.34 years (maximum days 7067) for ER negative, and 18.59 years (maximum days 6796) for ER positive groups and the survival analysis event was overall survival.

2.2 Chondrosarcoma Set's miRNA Expression Analysis

All file processing and data analyses were held in a Linux OS environment, primarily using the R functional language (R version 3.6.3 (2020-02-29)).

2.2.1 Data Downloading

Binarized versions (.sra extension) of raw .fastq sequencing result files were downloaded locally using the "prefetch" terminal command from SRA-toolkit (version X) and each Run's ID. These Reads were distributed in 102 folders. Then, with the SRA-toolkit's "fasterq-dump" command, these .sra files were converted to .fastq files, which is a human-readable and tool-parsable format. Specifically, in a .fastq file, each sequencing read consists of 4 lines. The first and third lines hold sequencing and metadata information, while the second and fourth lines store the generated sequence and the corresponding per-base quality scores, respectively.

2.2.2 Quality Control

For Illumina sequencing to take place, artificial sequences are attached-ligated to the selected biological fragments. These are PCR primers and adapter sequences, that enable each fragment to be amplified and attached to the Illumina flow cell, respectively. For accurate analysis, it is important to know if these artificial sequences have been sequenced along the biological fragments, and to remove them from the raw data. Additionally, an important step prior to analysis is the trimming of low-quality bases which usually appear at the 3' of reads. Quality control is essential because data would have fewer duplicates and won't have overrepresented sequences and overall data quality will be increased.

By using the FASTQC suite [68], a series of metrics that do quality control checks on raw sequence data coming from high throughput sequencing pipelines are taken into account by giving as an input the .sra.fastq. Then, by checking the resulting .html file, we are interested in seeing in the section "Overrepresented sequences". In the column "Possible source" we check if there is an adapter sequence in a relatively large percentage. In the chondrosarcoma dataset, no known adapter sequence was found. In order to identify the samples' adapter, "minion" command was used to search for the *3' adapter* without prior knowledge of its sequence. Minion expects FASTQ input, imports two million sequences by default and detects overrepresented sequences within the sequencing data. Minion's output shows the 10 most represented sequences. Over-represented sequences from both FASTQC and Minion were checked against miRBase so that no miRNA sequences would be selected for removal. Then, the remaining candidate sequences were tested using "grep" command in the .fastq files and the non-standard ACGGGCTAATATTTATCGGTGGAGCACTCACATCTC adapter sequence was revealed at the 3' of numerous reads.

2.2.3 Adapter Trimming

"Trim Galore" is a wrapper script to automate quality and adapter trimming. This tool trims the adapter sequence from the 3' end of all reads of the .fastq file and also can remove sequences if they become too short during the trimming process. By using "Trim galore" we obtained the final data that were used in this desertion.

2.2.4 Bowtie

After miRNA sequences' trimming, a genome index was used to align the trimmed miRNA sequences. Indices allow the aligner to narrow down the potential origin of a query sequence within the genome, saving both time and memory and it is proposed to use molecular indices for all RNA-Seq experiments [69]. Indexing computational strategies are essentially used to speed

up mapping algorithms. In this thesis, “Bowtie” was used, which is a software package commonly used in bioinformatics for sequence alignment and sequence analysis and it uses the Burrows-Wheeler transform algorithm. Bowtie is an ultrafast, memory-efficient aligner designed to quickly align large sets of short reads to large genomes and a genome index bowtie-build can index reference genomes of any size.

We created a genome index with the command “bowtie-build”. The genome index used was “GCA_000001405.15_GRCh38_no_alt_analysis_set” referring to the GRCh 38 human reference genome assembly that includes standard chromosomes and additional non-placed scaffolds. This index was during mapping-quantification by miRDeep (-p parameter in the “mapper” command of “mapper.pl”).

2.2.5 Commands example for the first Run:

---

prefetch ERR2820557 –output-directory . # ERR2820557 is the first Run’s ID

fasterq-dump -e 8 ERR2820557.sra -O . # 8 threads

fastqc ERR2820557.sra.fastq

minion search-adapter -k 12 -do 5000000 -show 10 -i ERR2820557.sra.fastq -o ERR2820557.sra.fastq.minion

trim_galore ERR2820557.sra.fastq -j 2 --quality 20 --phred33 –a ACGGGCTAATATTTATCGGTGGAGCACTCACATCTC --stringency 3 --length 16 --max_length 28 --trim-n --dont_gzip --fastqc --output_dir .

2.2.6 Mapper

“Mapper.pl” is a miRDeep2 script for miRNA preprocessing data sequencing. More specifically, “mapper.pl” processes reads and/or maps them to the reference genome. The Bowtie-made genome index was provided as reference genome, while clean, .fastq files, pre-processed by Trim-galore, were used as main inputs. The output of the mapper module can directly be used for identification of known and novel miRNAs in the data, and for their quantification[70].

The parameters used in this command were the following:
-e: input file is in fastq format
-i: mapping will be performed against genome and mapper.pl will convert rna to dna alphabet
-h: parsing will be to fasta format
-l: mapper.pl will discard reads shorter than the 16 nts
-m: same-sequence reads will be collapsed
-p: bowtie-build genome index full path

2.2.7 Quantifier

MiRNA's quantification is accomplished with "quantifier.pl" which is a miRDeep2 script designed to give an overview of expressed miRNAs in the data. As input, "quantifier.pl" takes the files generated by "mapper.pl". Also, files containing known miRBase mature and known miRBase precursor sequences in .fasta format have to be provided. These files can be downloaded from miRBase and then be unzipped with the "gunzip" command.

The parameters used in this command were the following:

-p: path pointing to miRBase-derived precursor sequences
-m: path pointing to miRBase-derived mature sequences
-r: files generated by "mapper.pl"
-d: suppress production of supplementary .pdf files during run
-W: read counts are weighed by their number of mappings
-t: species code (human species is "hsa")

2.2.8 Data Sequencing Script

Due to the large volume of data that was already being processed, a script was built to perform quantification, obtain metrics and prepare final quantification results in an automated manner. This R script contains of some functions that serve different purposes. Firstly, a function was created to run mapper and quantifier commands for all input files. Subsequently, a few functions were created in order to obtain metrics from .fastq files and from miRDeep2 results. Those functions' output is a table showing read numbers from raw .fastq files to quantification results. Another function was made to match quantification results with survival information, while the final function that was created built a count matrix by merging files from each quantification run. This count matrix is suitable for the application's input.

2.3 Application Development

2.3.1 Input Data Tables

The main survival analysis application was developed in R. Development took place in a Linux OS environment, however the application can work on any system running R (version 3.6) and the required packages, e.g. Windows or macOS. The main input are two data tables; one containing expression information and one containing time-to-event details. These inputs need to follow specific formats, as detailed below.

2.3.1.1 Count matrix

The count matrix must be a tab-delimited file containing expression values across all miRNAs and samples. Specifically, the first column (#miRNA) must contain unique miRNA names, while all following columns must contain the respective read counts from each individual dataset. The column names of the expression data must be unique sample identifiers. An example of the form of the count matrix is provides below in Figure 3.

	#miRNA	ERR2820557	ERR2820566	ERR2820656	ERR2820657	ERR2820658	ERR2820567	ERR2820568	ERR2820569
1	hsa-let-7a-2-3p	7.000	26.0000	26.50000	1.00000	8.5000	30.000	11.0000	11.000
2	hsa-let-7a-3p	443.000	97.0000	204.00000	1.00000	328.0000	260.000	87.0000	46.000
3	hsa-let-7a-5p	297072.000	155789.7214	329720.96429	2527.45000	139202.4000	162663.117	72904.2786	98880.350
4	hsa-let-7b-3p	680.000	679.5000	1729.00000	14.00000	169.5000	263.000	453.0000	518.000
5	hsa-let-7b-5p	33098.300	29680.1167	117799.40000	2108.90000	16887.8000	8639.133	42900.4000	30639.000
6	hsa-let-7c-3p	32.000	55.0000	15.00000	0.00000	84.0000	29.000	20.0000	35.000
7	hsa-let-7c-5p	14905.633	10533.0405	3186.52143	243.73333	11176.0167	5238.283	10452.9429	12542.417
8	hsa-let-7d-3p	135.000	189.0000	232.00000	36.00000	89.0000	43.000	215.0000	75.000
9	hsa-let-7d-5p	27026.500	22480.2500	28642.85000	503.25000	19533.0833	14562.333	10338.0000	6170.333
10	hsa-let-7e-3p	269.000	100.0000	192.50000	3.00000	210.5000	152.000	78.0000	57.000
11	hsa-let-7e-5p	7039.650	5076.0667	8495.88333	193.45000	5043.3833	5445.233	2964.2167	1722.533
12	hsa-let-7f-1-3p	124.000	66.5000	37.00000	0.00000	65.5000	94.000	23.0000	13.000
13	hsa-let-7f-2-3p	2.000	0.0000	1.00000	0.00000	3.0000	0.000	2.0000	0.000
14	hsa-let-7f-5p	276990.483	125483.9143	100373.50952	778.40000	241218.1000	228124.283	36560.7857	33048.033
15	hsa-let-7g-3p	1.000	1.0000	0.00000	0.00000	4.0000	2.000	1.0000	1.000
16	hsa-let-7g-5p	103482.600	53961.3238	64800.40476	502.06667	178431.6667	152511.050	23963.3262	21744.667
17	hsa-let-7i-3p	23.000	192.0000	59.00000	1.00000	307.0000	27.000	44.0000	18.000
18	hsa-let-7i-5p	83736.500	204400.5000	80643.50000	842.00000	336550.0000	167209.000	46681.5000	25306.833

Figure 3: The input count_matrix format

2.3.1.2 Survival table

The survival table is a three-column tab-delimited file. The first column, named “Run”, must contain unique experiment identifiers which must be the same as those provided in count matrix. The second column, “event_death”, is in binary format and informs if the event of interest has occurred or not for each patient/sample, with 0 meaning that the event did not happen within the monitoring time and 1 that it did. The last column, “overall_survival” is numeric and provides the elapsed time until the last follow-up (for cases where the event did not occur) or until the event occurrence (e.g. time of death). This information could be provided in any kind of time units (e.g. months, days). An example of the form of the survival table is given in the Figure 4 below.

	Run	event_death	overall_survival
1	ERR2820560	0	96
2	ERR2820561	1	10
3	ERR2820562	NA	NA
4	ERR2820563	0	20
5	ERR2820564	0	1
6	ERR2820565	1	3
7	ERR2820566	0	80
8	ERR2820567	1	7
9	ERR2820568	0	77
10	ERR2820569	0	26
11	ERR2820570	1	9
12	ERR2820571	0	38
13	ERR2820572	1	1
14	ERR2820573	0	78
15	ERR2820574	0	86
16	ERR2820575	0	89
17	ERR2820576	0	72
18	ERR2820577	0	0
19	ERR2820578	0	80

Figure 4: The input survival table format

2.3.2 R Package requirements

The main R package that was used in the application for fast and efficient table operations is "data.table". Some additional packages used in the application are "optparse" (a command line option parser), "edgeR" (a package for differential expression analysis; utilized to normalize the reads to CPM – counts per million units), "matrixStats" (to produce sample quantiles corresponding to the given probabilities), "survival" (for survival analysis) and "ggfortify" (to enable estimation of confidence intervals in survival curves). More accurately, the "survival" package utilizes the input data to perform a survival analysis and specifically Kaplan Meier curve analysis while the log rank test is performed.

2.3.3 Data Filtering

A script was developed and incorporated in the application, which implements a filtering strategy to retain only miRNAs with sufficiently large counts for statistical analysis. Specifically, it keeps miRNAs with read counts above "count.cutoff" in at least "sample.percentage" samples. We set "count.cutoff" at 4 reads and "sample.percentage" at 75% (0.75) of total samples.

2.3.4 Counts Per Million

CPM (Counts per Million Reads Mapped) value is a useful descriptive measure for the expression level of a gene, making up an equal transcriptome composition for all reads. In the occasion of having two RNA-Seq samples with the same number of total reads (sequencing depth), if one sample has a larger transcriptome, a smaller proportion of the reads will pertain to that feature. For samples with the same qualitative transcriptome and sequenced at the same depth, if a subset of other shared features is more highly expressed in one sample, it will make up a larger proportion of the total read pool, leaving a lower proportion of the reads mapping to the test feature. Thus, with CPM value we divide by the total mapped reads and then bring the units up to a more convenient number with the multiplication by one million. More specifically, for a variable r_i and for each feature i, CPM is the count of sequenced fragments mapping to the feature scaled by the total number of mapped reads (R) times one million.

$$CPM = \frac{r_i}{\frac{R}{10^6}} = \frac{r_i}{R}10^6$$

The CPM function in R is provided by the edgeR package.

2.3.5 Cutoff – Threshold

In order to define the high- and low-expression groups within the patient cohort, based on the expression of each miRNA, there is the need to define two thresholds, one upper and one lower. We use "quantile()" function of "matrixStats" package to generate the sample quantiles. A quantile *Q* is a point dividing the ranked set of *n* observations *i* (from smallest to largest) so that Q x n observations have value less that i_Q. Using a lower and an upper threshold (default thresholds 0.25 and 0.75 respectively), each patient is characterized as belonging to the low-expression group (encoded as 0), the high-expression group (encoded as 1) or in the grey zone in between (N/A value, not used in further analysis).

2.3.6 Optparse

"Optparse" package, which provides a command line option parser, was used in this application so that users can determine the following arguments directly through a terminal call: the count matrix (count_matrix), the survival information matrix (event_info), the output directory (output_dir) where application results should be written and the upper (up_threshold) and lower threshold (down_threshold) of the user's choice. The user can define all the aforementioned options using the arguments:

-c or --count_matrix, for "count_matrix", which accepts characters and has no default value
-e or --event_info, for "event_info", which accepts characters and has no default value
-o or --output_dir, for "output_dir", which accepts characters and uses the current directory as default
-u or --up_threshold, for "up_threshold", which accepts values from 0 to 1 (default value 0.75)
-d or --down_threshold for "down_threshold", which accepts values from 0 to 1 (default value 0.25)

2.3.7 Kaplan Meier and Log Rank Test

We were interested in comparing the survival times between two populations. The Kaplan-Meier method is the most common way to estimate survival times and probabilities. It is a non-parametric approach, independently described by Edward Kaplan and Paul Meier and conjointly published in 1958 in the Journal of the American Statistical Association, that results in a step function, where there is a step down each time an event occurs.

We used "survdiff()" function from "survival" R package that tests statistical difference between the survival times. "Survdiff()" needs a survival-type object as an argument, created using "Surv()" function, also from "survival" R package. The first argument of the survival object is an event vector, the second argument is time (event occurrence or last follow-up) and an indicator vector for right-censoring is input as the third argument. Converted to our application's vectors, survival object's first argument is "overall_survival", the second argument is "event_death" and the third is "type". Basically, the event vector declares if the event happened for each patient, the time vector indicates the elapsed time and the indicator vector informs of the group (high miRNA expression or low) each patient belongs to.

In order to compare survival curves of the two groups, the log-rank test is applied, which is a statistical hypothesis test that tests the null hypothesis (H_0) that survival curves of two populations do not differ. The two hypotheses are:

H_0: $h_1(t) = h_2(t) = \ldots = h_n(t)$, means that survival curves of two populations do not differ
H_1: $h_i(t_0) \neq h_j(t_0)$, means that survival curves of two populations differ

h_n : hypothesis
t : observation time
t_i : times where the event was observed

A certain probability distribution, namely a Chi-squared distribution, can be used to derive a p-value. The Chi-squared distribution is a statistical method that is used to determine if two categorical variables have a significant correlation between them. P-values are used in statistical hypothesis testing to quantify statistical significance.

We used “pchisq()” function, which is a distribution function and returns the probability that a variable that follows the Chi-square distribution is smaller or equal to diff$chisq. The “diff$chisq” variable means that we selected every column in “diff” variable and used the Chi-squared distribution. Function’s “pchisq()” result is the p-value which determines if a specific miRNA expression is statistically significant based on the survival event (patient’s death) and the event’s censored time.

A result with $p < 0.05$ is usually considered significant. In this application when $p < 0.05$ a Kaplan Meier plot is made with the “autoplot” function. “autoplot()” function’s argument is a survfit object from “survfit()” function. “Survfit()” function computes an estimate of a survival curve for censored data using the Kaplan-Meier method.

2.3.8 The central application function “run_surv”

The main function is “run_surv” and its arguments are:

i. onemir: one row of the survival table, containing read count data for one miRNA from all patients (samples)
ii. mirna_name: the name of the miRNA that the function is running for
iii. up_threshold: user-defined upper threshold
iv. down_threshold: user-defined lower threshold

The “run_surv” function uses the Cutoff – Threshold method (2.3.5) to categorize each sample/patient into “high” and “low” groups, based on the expression of one miRNA at a time. An example of the “surv_table” is in Figure 5:

	Run	event_death	overall_survival	type
1	ERR2820561	1	10	1
2	ERR2820562	NA	NA	1
3	ERR2820563	0	20	1
4	ERR2820572	1	1	1
5	ERR2820573	0	78	1
6	ERR2820574	0	86	1
7	ERR2820575	0	89	1
8	ERR2820577	0	0	1
9	ERR2820578	0	80	1
10	ERR2820579	1	95	0
11	ERR2820581	NA	NA	0
12	ERR2820582	0	27	0
13	ERR2820585	0	62	0
14	ERR2820586	0	28	0
15	ERR2820588	0	20	0
16	ERR2820590	1	19	1
17	ERR2820591	1	6	1
18	ERR2820559	0	126	0
19	ERR2820557	0	41	0

Figure 5: ‘Surv_table”

Then, in a variable called “classif_txt” we store the full path that the application saved the individual for every miRNA files that contain the SRA IDs and the miRNA expression analysis “type”. After that, the log rank test and the Kaplan Meier method is applied (2.3.7) in order to

store, in the file “Kaplan_Meier_Plots”, the plots that occurred after the p-value was statistically significant. If the p-value was statistically significant, the “plot path" is determined, the plot is being created and saved to the above-mentioned “plot_path”. If the p-value is not statistically significant the plot path and the “classif_txt” variable get the value N/A.

A data table is made with columns “mirna_name”, “plot_path” and “classif_txt” and stored in the “my_vector”, which will be returned after the “run_surv” function.

2.3.9 Running the application

The main application can run in the command line (terminal) of a Linux OS as shown below:

```
Rscript KM_survival_opt.R -u .75 -d .25 -c counts.txt -e survival.txt -o
output_folder
```

With:
KM_survival.R : being the application’s name
-u .75: the upper threshold chosen is 0.75
-d .25: the lower threshold chosen is 0.25
-c counts.txt: being the count matrix in .txt format
-e survival.txt: being the survival table in .txt format

The application starts with parsing into R objects the user-provided arguments. Then folders miRNAs_classification and Kaplan_Meier_Plots are created in case they do not exist in the provided output directory. After that, the Data Filtering (2.3.3) and Counts Per Million (2.3.4) operations were applied to count matrix.

The “run_surv” (2.2.8) function is executed consecutively with “lapply” for all miRNAs that pass the filtering criteria. The “lapply” function applies a function of our choice, in this case “run_surv”, over a list of vectors. “lapply’s” function results which is the “my_ vector” (2.2.8).

Finally, the application sorts the lapply() results based on the lowest p-value and saves this result to selected output path.

2.3.10 Output

The application’s output is written in three folders: “/Kaplan_Meier_Plots/”, “/miRNAs_classification/” and “/survival_results/”.

1. **/Kaplan_Meier_Plots/** contains Kaplan Meier plots, in TIFF graphics format, for every miRNA that was considered to be statistically significant
2. **/miRNAs_classification/** contains tab-delimited files providing the classification of each sample into the low- or high-expression group, for each tested miRNA
3. **/survival_results/** contains a tab-delimited file showing for each tested miRNA, the respective log rank test p-value, the full path of the Kaplan Meier Plot (if applicable) and a full path towards the cohort dichotomization information

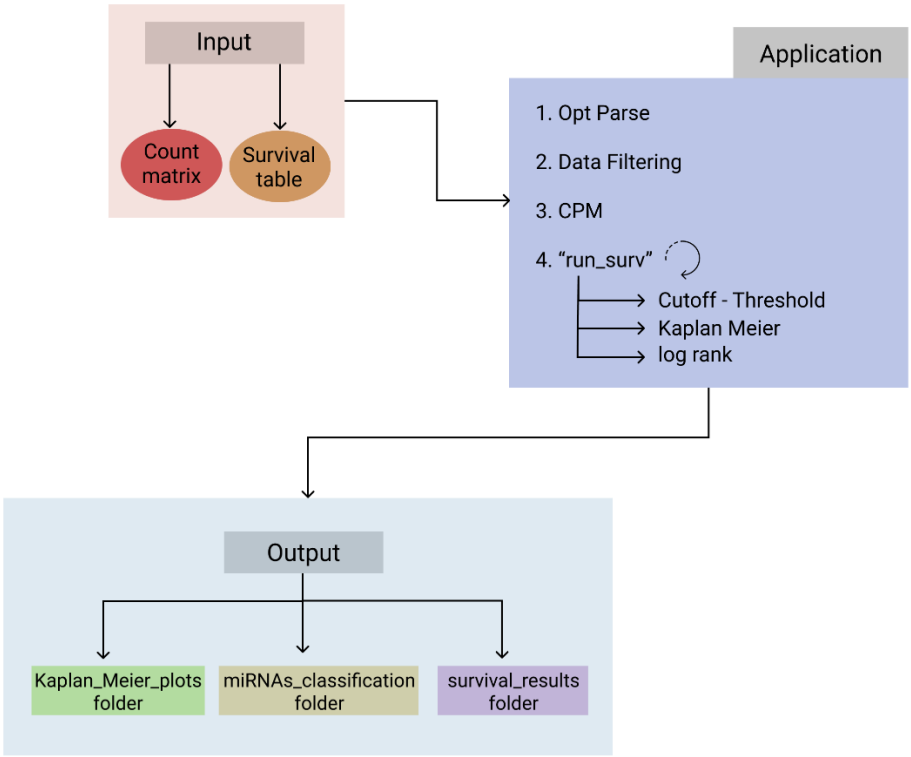

Figure 6: The application's workflow

3 Results and Discussion

3.1 Output Results

During the application development, measurable results were produced (selected thresholds: upper threshold = 0.75, lower threshold = 0.25) and are discussed below. Since we had run individually for different datasets, we discuss firstly for each dataset one by one. In each dataset's results we have also annotated whether low or high expression of each miRNA is positively associated with survival.

3.1.1 Chondrosarcoma

We had 196 significant miRNAs for the Chondrosarcoma dataset, that means that 196 miRNAs had p-value less than 0.05 and their log rank test was statistically significant. Below there is the table (Table 1) with the top 10 most statistically significant miRNAs for the Chondrosarcoma dataset. For 8 of the most significant miRNAs, low expression is in favor of survival, while for 2 the opposite is true.

Table 1: 10 most statistically significant miRNAs in Chondrosarcoma dataset

miRNA	Expression in Favor of Survival	p-value
1. hsa-miR-17-5p	Low	7,13E+08
2. hsa-miR-664a-5p	High	8,36E+08
3. hsa-miR-218-5p	Low	1,73E+09
4. hsa-miR-130b-3p	Low	1,97E+08
5. hsa-miR-20a-5p	Low	2,72E+09
6. hsa-miR-93-5p	Low	5,53E+09
7. hsa-miR-140-3p	Low	5,60E+09
8. hsa-miR-23b-5p	High	0.00011
9. hsa-miR-660-5p	Low	0.00013
10. hsa-miR-501-5p	Low	0.00014

The Kaplan Meier plot for hsa-miR-17-5p is provided in Figure 7. As shown, patients with lower expression of this miRNA had better survival prognosis. Studies in hepatocellular carcinoma and pancreatic cancer have found that high expression of miR-17-5p is negatively correlated with overall survival and disease-free survival [71-73], therefore our results extend this finding also in Chondrosarcoma.

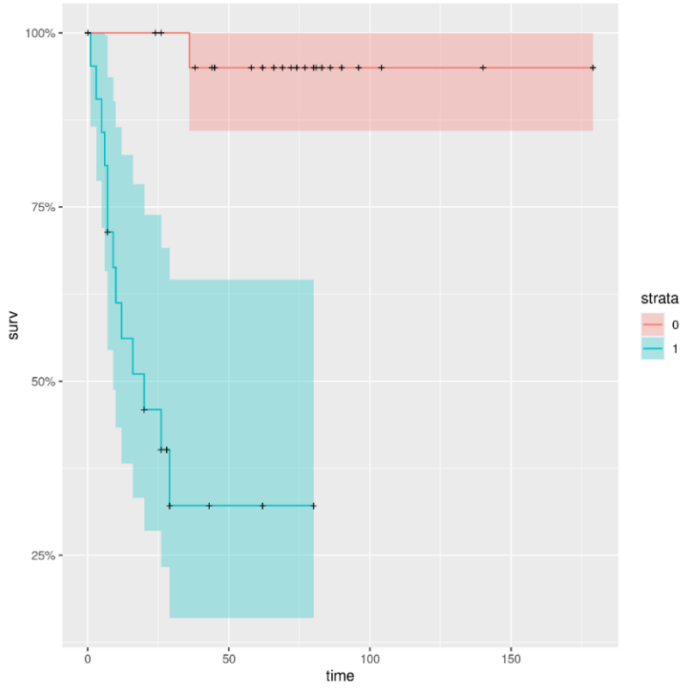

Figure 7: Kaplan Meier curves of miR-17-5p in Chondrosarcoma. Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

The Kaplan Meier curve for miRNA hsa-miR-664a-5p is the Figure 8. This curve shows that higher expression of hsa-miR-664a-5p is associated with a chance of better survival.

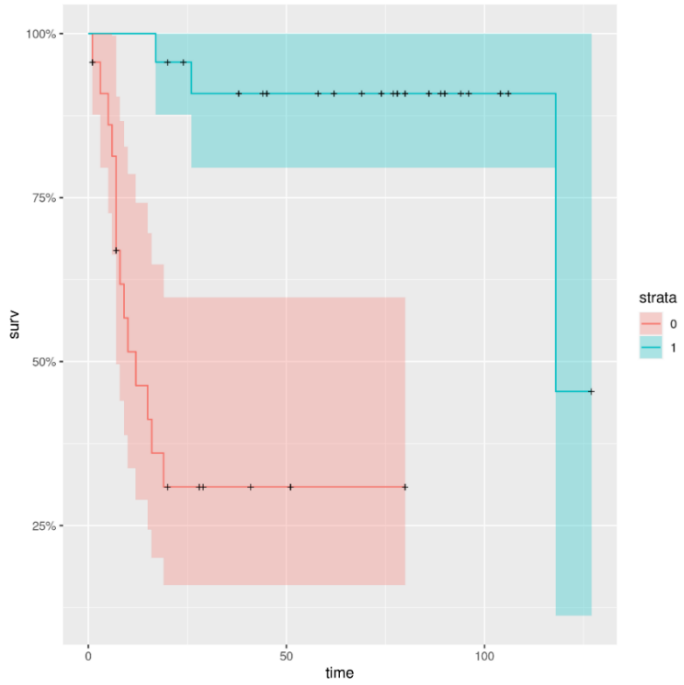

Figure 8: Kaplan Meier curves of hsa-miR-664a-5p in Chondrosarcoma, Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

3.1.2 TCGA – Breast Cancer: ER-

We had 11 significant miRNAs for the TCGA dataset of ER- breast cancer, that means that 11 miRNAs had p-value less than 0.05 and their log rank test was statistically significant 11 miRNAs that passed the log rank p-value threshold and were worthy of further statistical observation. Below we can see the Table 2 with their p-value and if the higher or the lower expression of them is in favor of survival prognosis. 8 of the miRNAs are highly expressed and 3 of them are expressed in lower quantities.

Table 2: 11 most statistically significant miRNAs in TCGA-BRC ER negative dataset

miRNA	Expression in Favor of Survival	p-value
1. hsa-miR-29c-3p	high	0.00390
2. hsa-miR-342-5p	high	0.00663
3. hsa-miR-148b-3p	low	0.01377
4. hsa-let-7i-3p	high	0.01973
5. hsa-miR-625-3p	high	0.02160
6. hsa-miR-320b	high	0.03122
7. hsa-miR-509-3p	high	0.03283
8. hsa-miR-629-3p	high	0.03615
9. hsa-let-7f-1-3p	low	0.04207
10. hsa-miR-370-3p	high	0.04688
11. hsa-miR-24-3p	low	0.04998

miRNA's hsa-miR-29c-3p Kaplan Meier curve is presented below and it could be interpreted with the following sentences. We could predict that an ER- breast cancer patient could have better chance of surviving if hsa-miR-29c-3p is found in higher quantities. Also, we could observe that in this plot the confidence intervals for the higher and lower miRNA quantity are overlapping, leading us to be unsure if hsa-miR-29c-3p could be a possible biomarker. There are no studies directly confirming or rejecting our application's results when it comes to hsa-miR-29c-3p. There is a study mentioning that the specific miRNA's expression is decreased along disease progression of breast cancers, however, it is not specific on 3p or 5p and there is no mention of ER status consideration [74]. Another study on miR-29c focuses on its therapeutic role and not on the one as a prognostic biomarker [75].

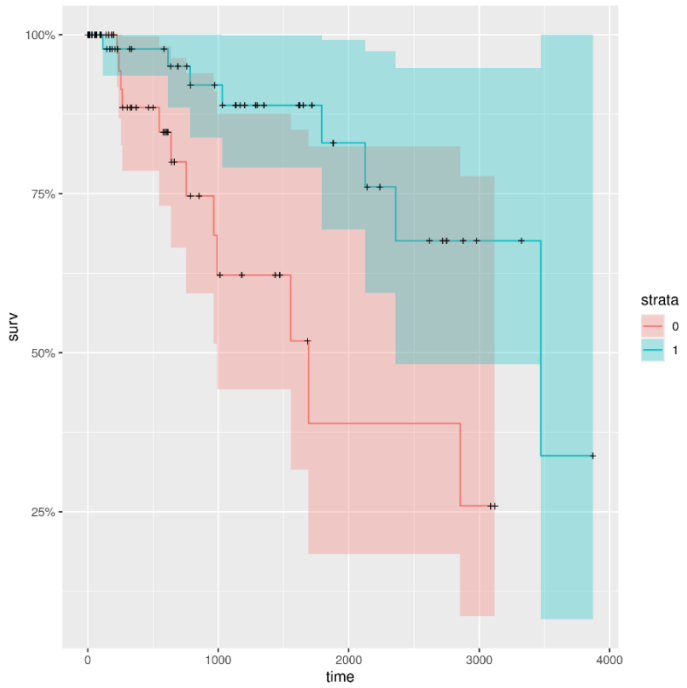

Figure 9: Kaplan Meier curves of hsa-miR-29c-3p in Breast Cancer, ER- subtype, Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

An example of a miRNA, where its lower quantity is associated with a better survival chance in ER- breast cancer is hsa-miR-148b-3p. Below in Figure 10, we can observe its Kaplan Meier curves. Again, the confidence intervals are overlapping, though in this case, we cannot say from the overlapping alone that this miRNA could not be a possible biomarker.

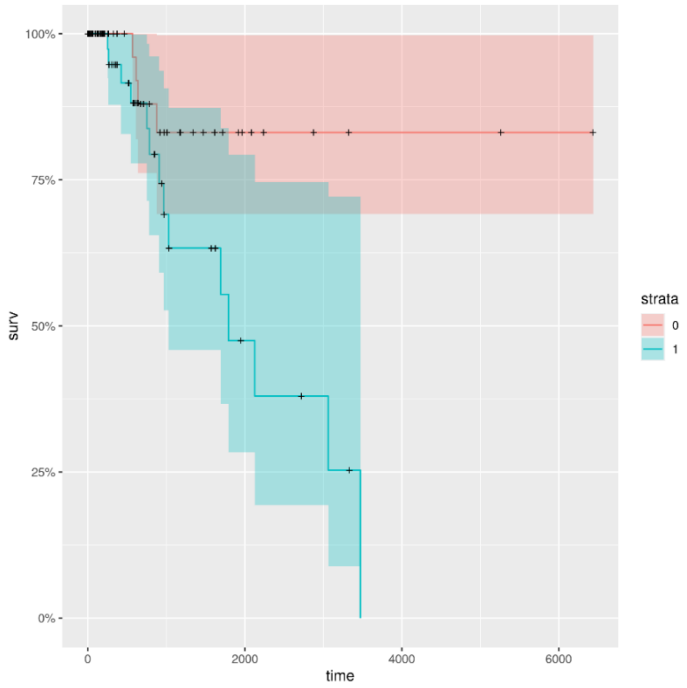

Figure 10: Kaplan Meier curves of hsa-miR-148b-3p in Breast Cancer, ER- subtype, Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

3.1.3 TCGA – Breast Cancer: ER+

There were 94 significant miRNAs for the TCGA dataset of breast cancer with ER+, that means that 94 miRNAs had p-value less than 0.05 and their log rank test was statistically significant. In Table 3, there are 10 miRNAs with the higher p-value score.

Table 3: 10 most statistically significant miRNAs in TCGA-BRC ER positive dataset

miRNA	Expression in Favor of Survival	p-value
1. hsa-miR-874-3p	Low	4,83E+07
2. hsa-miR-6511a-3p	Low	2,50E+08
3. hsa-miR-328-3p	Low	2,09E+09
4. hsa-miR-3127-5p	Low	3,30E+09
5. hsa-miR-484	Low	4,55E+09
6. hsa-miR-574-5p	Low	4,85E+09
7. hsa-miR-99b-5p	Low	7,05E+09
8. hsa-miR-99b-3p	Low	0.00010
9. hsa-miR-4510	High	0.00014
10. hsa-miR-125a-5p	Low	0.00017

9 of the most significant miRNAs are found in lower quantities could possibly predict a better survival expectancy and 1 of them in higher quantities.

miRNA's hsa-miR-874-3p Kaplan Meier curve is the one stated below. It shows that patients with lower expression of this miRNA could have a better chance of surviving. But the confidence intervals are overlapping at some point, so it is unsure if hsa-miR-874-3p could be a robust biomarker. A study found that the expression level of miR-874 is down-regulated in breast cancer in comparison with adjacent normal tissue [76]. This study indicates that miR-874 is probably higher in normal tissue; however, our findings are that among patient tissue in ER+, higher means worse survival.

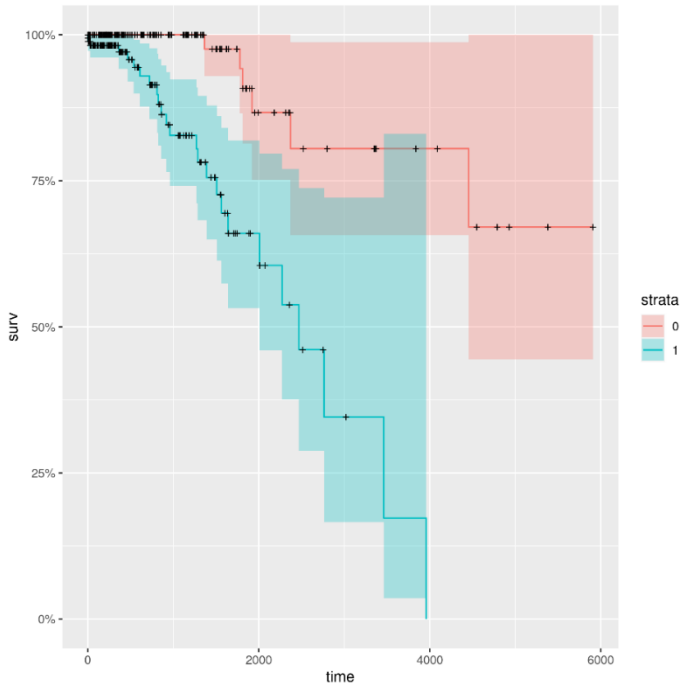

Figure 11: Kaplan Meier curves of hsa-miR-874-3p in Breast Cancer, ER+ subtype, Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

On the contrary, when miRNA hsa-miR-4510 is found in higher quantities the patients that exhibit better survival. MiRNA's hsa-miR-4510 Kaplan Meier curve is shown in Figure 12.

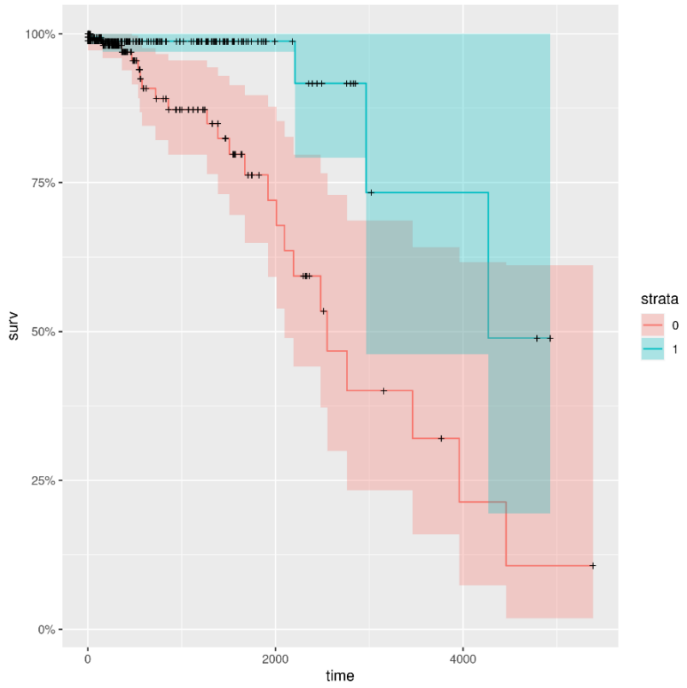

Figure 12: Kaplan Meier curves of hsa-miR-4510 in Breast Cancer, ER+ subtype, Stratum 0: patient group that expresses low quantities of miRNA of interest. Stratum 1: patient group that expresses high quantities of miRNA of interest

3.1.4 Comparison of findings between ER+ and ER- breast cancer patients

In the two TCGA-BRCA datasets, there are some miRNAs that are common. In the following Venn Diagram, Figure 13, we could see that from the 11 statistically significant miRNAs in the breast cancer with ER- dataset and from the 94 most statistically significant miRNAs in the breast cancer with ER+ dataset there are 2 common miRNAs: hsa-miR-509-3p and hsa-miR-148b-3p.

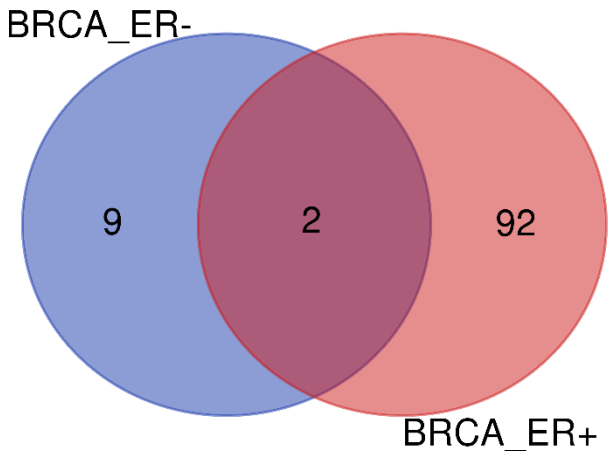

Figure 13: Venn Diagram of the application's results with breast cancer with ER- dataset and with breast cancer with ER+ dataset

In Table 4, the p-values of the two common miRNAs in the two datasets are shown.

Table 4: The common miRNAs from the application's results with breast cancer with ER- dataset and with breast cancer with ER+ dataset

miRNA	Expression in Favor of Survival	p-value	Dataset
hsa-miR-148b-3p	Low	0.01377	Breast Cancer: ER-
hsa-miR-148b-3p	Low	0.00352	Breast Cancer: ER+
hsa-miR-509-3p	High	0.03283	Breast Cancer: ER-
hsa-miR-509-3p	Low	0.01004	Breast Cancer: ER+

miR-509-3p depending on the ER status, changes its correlation with survival positively or negatively. More specifically, we observe that when miR-509-3p is highly expressed in breast cancer patients with ER- there could be a lower risk of them passing away. On the contrary, in breast cancer patients with ER+ if miR-509-3p is in lower quantities, the patient is more likely to survive. There is a possibility that the ER status plays a crucial role for this miRNA. to our knowledge, no study specifically assessing the association of ER status with miR-509-3p expression and the potential mechanisms involved.

Previously, it was reported that miR-509-3p functions as a tumor suppressor, potentially serving prominent roles in the development of various types of cancer, including breast cancer [77]. However, although there is a large body of research, the functions of miR-509-3p require further investigation [78].

When miR-148b-3p is expressed in lower quantities in breast cancer patients it could lead to better survival prognosis in both ER+ and ER- status. That could possibly mean that miR-148b-3p could be used for survival analysis of both ER positive and negative status. The prognostic role of miR-148-3p breast cancer tissue expression regardless of ER status has also been confirmed before by Dai *et al.* [79]. Interestingly, in another study, miR-148b-3p serum levels were found significantly lower in breast cancer patients than in disease-free controls. In other words, the opposite expression of this miRNA in serum is a potential diagnostic biomarker for breast cancer.

4 Conclusions

Within this thesis, an application for survival analysis utilizing the abundance of miRNAs was developed. It uses a sRNA-Seq count matrix and an event information table to perform dichotomization of a cohort of interest based on miRNA expression and to provide as output Kaplan Meier curves and statistics metrics of the dichotomization significance. This application is easy to use and could be also used for other data types (e.g. gene expression count matrix), without or with minimal changes to its code. The application was tested using overall survival as an event of interest, however any other event that is observed in the course of time could also be used instead.

A number of limitations in our application currently exist. From a technical standpoint, the developed scripts have not been set to perform error handling and output informative messages to users if they provide wrong arguments. For example, the application should check the validity of the given output path and place a "/" if needed. Also, the formatting of the input tables is not checked for consistency when they are loaded as objects in R environment.
Furthermore, a number of changes could be made in the main function of the application. We could enable users to also perform univariate or multivariate Cox proportional hazards regression. Cox regression is a method for investigating the effect of several variables upon the time a specified event happens, and could be used to inspect miRNA expression together with other variables, such as age. Also, a method to inspect the 95% confidence intervals of the produced Kaplan Meier curves and report instances that overlap the less would be a useful addition to our application.

miRNAs, as well as many other indicators, are studied on a large scale, with robust high throughput analysis methods to examine their role as potential prognostic biomarkers. This application was developed and tested on miRNAs using cancer tissue samples to inspect the miRNAs function as possible overall survival indicators. As a case study, our application was used to highlight a number of miRNAs with potential prognostic roles in Chondrosarcoma, ER+ and ER- breast cancer.

References

[1] R.C. Lee, V. Ambros, An extensive class of small RNAs in Caenorhabditis elegans, Science 294(5543) (2001) 862-864.
[2] S.M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, T. Tuschl, Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells, Nature 411(6836) (2001) 494-498.
[3] V. Ambros, R.C. Lee, A. Lavanway, P.T. Williams, D. Jewell, MicroRNAs and other tiny endogenous RNAs in C-elegans, Current Biology 13(10) (2003) 807-818.
[4] T. Kuwabara, J. Hsieh, K. Nakashima, K. Taira, F.H. Gage, A small modulatory dsRNA specifies the fate of adult neural stem cells, Cell 116(6) (2004) 779-793.
[5] V. Ambros, B. Bartel, D.P. Bartel, C.B. Burge, J.C. Carrington, X.M. Chen, G. Dreyfuss, S.R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, T. Tuschl, A uniform system for microRNA annotation, Rna 9(3) (2003) 277-279.
[6] V.N. Kim, MicroRNA biogenesis: Coordinated cropping and dicing, Nature Reviews Molecular Cell Biology 6(5) (2005) 376-385.
[7] V.N. Kim, Small RNAs: Classification, biogenesis, and function, Molecules and Cells 19(1) (2005) 1-15.
[8] D.P. Bartel, Metazoan MicroRNAs, Cell 173(1) (2018) 20-51.
[9] D.P. Bartel, MicroRNAs: Target Recognition and Regulatory Functions, Cell 136(2) (2009) 215-233.
[10] J.M. Friedman, P.A. Jones, MicroRNAs: critical mediators of differentiation, development and disease, Swiss Medical Weekly 139(33-34) (2009) 466-472.
[11] R. Garzon, G.A. Calin, C.M. Croce, MicroRNAs in Cancer, Annual Review of Medicine 60 (2009) 167-179.
[12] V. Ambros, MicroRNAs and developmental timing, Current Opinion in Genetics & Development 21(4) (2011) 511-517.
[13] J. Starega-Roslan, E. Koscianska, P. Kozlowski, W.J. Krzyzosiak, The role of the precursor structure in the biogenesis of microRNA, Cellular and Molecular Life Sciences 68(17) (2011) 2859-2871.
[14] R. Iuliano, M.F.M. Vismara, V. Dattilo, F. Trapasso, F. Baudi, N. Perrotti, The Role of MicroRNAs in Cancer Susceptibility, Biomed Research International 2013 (2013).
[15] R.C. Lee, R.L. Feinbaum, V. Ambros, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75(5) (1993) 843-54.
[16] B. Wightman, I. Ha, G. Ruvkun, POSTTRANSCRIPTIONAL REGULATION OF THE HETEROCHRONIC GENE LIN-14 BY LIN-4 MEDIATES TEMPORAL PATTERN-FORMATION IN C-ELEGANS, Cell 75(5) (1993) 855-862.
[17] B.J. Reinhart, F.J. Slack, M. Basson, A.E. Pasquinelli, J.C. Bettinger, A.E. Rougvie, H.R. Horvitz, G. Ruvkun, The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans, Nature 403(6772) (2000) 901-906.
[18] V. Ambros, H.R. Horvitz, HETEROCHRONIC MUTANTS OF THE NEMATODE CAENORHABDITIS-ELEGANS, Science 226(4673) (1984) 409-416.
[19] M.Z. Michael, S.M. O'Connor, N.G.V. Pellekaan, G.P. Young, R.J. James, Reduced accumulation of specific microRNAs in colorectal neoplasia, Molecular Cancer Research 1(12) (2003) 882-891.
[20] G.A. Calin, C.D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, C.M. Croce, Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic

leukemia, Proceedings of the National Academy of Sciences of the United States of America 99(24) (2002) 15524-15529.
[21] P.S. Eis, W. Tam, L.P. Sun, A. Chadburn, Z.D. Li, M.F. Gomez, E. Lund, J.E. Dahlberg, Accumulation of miR-155 and BIC RNA in human B cell lymphomas, Proceedings of the National Academy of Sciences of the United States of America 102(10) (2005) 3627-3632.
[22] S.M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K.L. Reinert, D. Brown, F.J. Slack, RAS is regulated by the let-7 MicroRNA family, Cell 120(5) (2005) 635-647.
[23] D.E. Giza, G.A. Calin, microRNA and Chronic Lymphocytic Leukemia, Adv Exp Med Biol 889 (2015) 23-40.
[24] B.P. Lewis, C.B. Burge, D.P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, Cell 120(1) (2005) 15-20.
[25] V.N. Kim, J. Han, M.C. Siomi, Biogenesis of small RNAs in animals, Nature Reviews Molecular Cell Biology 10(2) (2009) 126-139.
[26] M.W. Jones-Rhoades, D.P. Bartel, B. Bartel, MicroRNAs and their regulatory roles in plants, Annual Review of Plant Biology 57 (2006) 19-53.
[27] Y. Lee, M. Kim, J.J. Han, K.H. Yeom, S. Lee, S.H. Baek, V.N. Kim, MicroRNA genes are transcribed by RNA polymerase II, Embo Journal 23(20) (2004) 4051-4060.
[28] G.M. Borchert, W. Lanier, B.L. Davidson, RNA polymerase III transcribes human microRNAs, Nature Structural & Molecular Biology 13(12) (2006) 1097-1101.
[29] M. Faller, F. Guo, MicroRNA biogenesis: there's more than one way to skin a cat, Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms 1779(11) (2008) 663-667.
[30] S.L. Lin, D. Chang, D.Y. Wu, S.Y. Ying, A novel RNA splicing-mediated gene silencing mechanism potential for genome evolution, Biochemical and Biophysical Research Communications 310(3) (2003) 754-760.
[31] X.Z. Cai, C.H. Hagedorn, B.R. Cullen, Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs, Rna 10(12) (2004) 1957-1966.
[32] R.I. Gregory, K.P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, R. Shiekhattar, The Microprocessor complex mediates the genesis of microRNAs, Nature 432(7014) (2004) 235-240.
[33] J.J. Han, Y. Lee, K.H. Yeom, Y.K. Kim, H. Jin, V.N. Kim, The Drosha-DGCR8 complex in primary microRNA processing, Genes & Development 18(24) (2004) 3016-3027.
[34] E. Lund, S. Guttinger, A. Calado, J.E. Dahlberg, U. Kutay, Nuclear export of microRNA precursors, Science 303(5654) (2004) 95-98.
[35] P.D. Zamore, T. Tuschl, P.A. Sharp, D.P. Bartel, RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals, Cell 101(1) (2000) 25-33.
[36] A. Vermeulen, L. Behlen, A. Reynolds, A. Wolfson, W.S. Marshall, J. Karpilow, A. Khvorova, The contributions of dsRNA structure to Dicer specificity and efficiency, Rna 11(5) (2005) 674-682.
[37] B.R. Cullen, Viruses and microRNAs, Nature Genetics 38 (2006) S25-S30.
[38] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, T. Tuschl, Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs, Molecular Cell 15(2) (2004) 185-197.
[39] A.E. Pasquinelli, S. Hunter, J. Bracht, MicroRNAs: a developing story, Current Opinion in Genetics & Development 15(2) (2005) 200-205.
[40] D.T. Humphreys, B.J. Westman, D.I.K. Martin, T. Preiss, MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function, Proceedings of the National Academy of Sciences of the United States of America 102(47) (2005) 16961-16966.

[41] R.S. Pillai, S.N. Bhattacharyya, C.G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, W. Filipowicz, Inhibition of translational initiation by Let-7 microRNA in human cells, Science 309(5740) (2005) 1573-1576.
[42] S.N. Bhattacharyya, R. Habermacher, U. Martine, E.I. Closs, W. Filipowicz, Relief of microRNA-mediated translational repression in human cells subjected to stress, Cell 125(6) (2006) 1111-1124.
[43] P.A. Maroney, Y. Yu, J. Fisher, T.W. Nilsen, Evidence that microRNAs are associated with translating messenger RNAs in human cells, Nature Structural & Molecular Biology 13(12) (2006) 1102-1107.
[44] C.P. Petersen, M.E. Bordeleau, J. Pelletier, P.A. Sharp, Short RNAs repress translation after initiation in mammalian cells, Molecular Cell 21(4) (2006) 533-542.
[45] R.F. Place, L.C. Li, D. Pookot, E.J. Noonan, R. Dahiya, MicroRNA-373 induces expression of genes with complementary promoter sequences, Proceedings of the National Academy of Sciences of the United States of America 105(5) (2008) 1608-1613.
[46] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebet, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H.R. Horvitz, T.R. Golub, MicroRNA expression profiles classify human cancers, Nature 435(7043) (2005) 834-838.
[47] J.F. Chen, E.P. Murchison, R. Tang, T.E. Callis, M. Tatsuguchi, Z. Deng, M. Rojas, S.M. Hammond, M.D. Schneider, C.H. Selzman, G. Meissner, C. Patterson, G.J. Hannon, D.Z. Wang, Targeted deletion of Dicer in the heart leads to dilated cardiomyopathy and heart failure, Proceedings of the National Academy of Sciences of the United States of America 105(6) (2008) 2111-2116.
[48] J. Kocerha, S. Kauppinen, C. Wahlestedt, microRNAs in CNS Disorders, Neuromolecular Medicine 11(3) (2009) 162-172.
[49] X.Q. Tang, G.L. Tang, S. Ozcan, Role of microRNAs in diabetes, Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms 1779(11) (2008) 697-701.
[50] M.N. Poy, M. Spranger, M. Stoffel, microRNAs and the regulation of glucose and lipid metabolism, Diabetes Obesity & Metabolism 9 (2007) 67-73.
[51] R. Grassmann, K.T. Jeang, The roles of microRNAs in mammalian virus infection, Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms 1779(11) (2008) 706-711.
[52] S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, A.J. Enright, miRBase: microRNA sequences, targets and gene nomenclature, Nucleic Acids Research 34 (2006) D140-D144.
[53] M.R. Friedlander, S.D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, Nucleic Acids Research 40(1) (2012) 37-52.
[54] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biology 10(3) (2009).
[55] M.K. Goel, P. Khanna, J. Kishore, Understanding survival analysis: Kaplan-Meier estimate, Int J Ayurveda Res 1(4) (2010) 274-8.
[56] M.A. Macha, P. Seshacharyulu, S.R. Krishn, P. Pai, S. Rachagani, M. Jain, S.K. Batra, MicroRNAs (miRNAs) as Biomarker(s) for Prognosis and Diagnosis of Gastrointestinal (GI) Cancers, Current Pharmaceutical Design 20(33) (2014) 5287-5297.
[57] P.S. Mitchell, R.K. Parkin, E.M. Kroh, B.R. Fritz, S.K. Wyman, E.L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K.C. O'Briant, A. Allen, D.W. Lin, N. Urban, C.W. Drescher, B.S. Knudsen, D.L. Stirewalt, R. Gentleman, R.L. Vessella, P.S. Nelson, D.B. Martin, M. Tewari, Circulating microRNAs as stable blood-based markers for cancer detection, Proceedings of the National Academy of Sciences of the United States of America 105(30) (2008) 10513-10518.

[58] A. Turchinovich, L. Weiz, A. Langheinz, B. Burwinkel, Characterization of extracellular circulating microRNA, Nucleic Acids Research 39(16) (2011) 7223-7233.
[59] A. Turchinovich, L. Weiz, B. Burwinkel, Extracellular miRNAs: the mystery of their origin and function, Trends in Biochemical Sciences 37(11) (2012) 460-465.
[60] X. Chen, Y. Ba, L.J. Ma, X. Cai, Y. Yin, K.H. Wang, J.G. Guo, Y.J. Zhang, J.N. Chen, X. Guo, Q.B. Li, X.Y. Li, W.J. Wang, Y. Zhang, J. Wang, X.Y. Jiang, Y. Xiang, C. Xu, P.P. Zheng, J.B. Zhang, R.Q. Li, H.J. Zhang, X.B. Shang, T. Gong, G. Ning, K. Zen, J.F. Zhang, C.Y. Zhang, Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases, Cell Research 18(10) (2008) 997-1006.
[61] X.X. Zeng, X. Zhang, Q. Zou, Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks, Briefings in Bioinformatics 17(2) (2016) 193-203.
[62] X. Chen, D. Xie, Q. Zhao, Z.H. You, MicroRNAs and complex diseases: from experimental results to computational models, Briefings in Bioinformatics 20(2) (2019) 515-539.
[63] A. Baldassarre, C. Felli, G. Prantera, A. Masotti, Circulating microRNAs and Bioinformatics Tools to Discover Novel Diagnostic Biomarkers of Pediatric Diseases, Genes 8(9) (2017).
[64] S. Tastsoglou, M. Miliotis, I. Kavakiotis, A. Alexiou, E.C. Gkotsi, A. Lambropoulou, V. Lygnos, V. Kotsira, V. Maroulis, D. Zisis, G. Skoufos, A.G. Hatzigeorgiou, PlasmiR: A Manual Collection of Circulating microRNAs of Prognostic and Diagnostic Value, Cancers 13(15) (2021).
[65] F. Russo, S. Di Bella, G. Nigita, V. Macca, A. Lagana, R. Giugno, A. Pulvirenti, A. Ferro, miRandola: Extracellular Circulating MicroRNAs Database, Plos One 7(10) (2012).
[66] R. Nicolle, M. Ayadi, A. Gomez-Brouchet, L. Armenoult, G. Banneau, N. Elarouci, M. Tallegas, A.V. Decouvelaere, S. Aubert, F. Redini, B. Marie, C. Labit-Bouvier, N. Reina, M. Karanian, L.R. Le Nail, P. Anract, F. Gouin, F. Larousserie, A. de Reynies, G. de Pinieux, Integrated molecular characterization of chondrosarcoma reveals critical determinants of disease progression, Nature Communications 10 (2019).
[67] A. Alexiou, D. Zisis, I. Kavakiotis, M. Miliotis, A. Koussounadis, D. Karagkouni, A.G. Hatzigeorgiou, DIANA-mAP: Analyzing miRNA from Raw NGS Data to Quantification, Genes 12(1) (2021).
[68] S. Andrews, FastQC: a quality control tool for high throughput sequence data., 2010.
[69] T. MM, R. J., A Molecular Indexing for Improved RNA-Seq., *Journal of Biomolecular Techniques : JBT*, 2014.
[70] S.D. Mackowiak, Identification of novel and known miRNAs in deep-sequencing data with miRDeep2, Curr Protoc Bioinformatics Chapter 12 (2011) Unit 12.10.
[71] L. Chen, M. Jiang, W.J. Yuan, H.H. Tang, miR-17-5p as a Novel Prognostic Marker for Hepatocellular Carcinoma, Journal of Investigative Surgery 25(3) (2012) 156-161.
[72] J.J. Zheng, P.H. Dong, S.M. Gao, N. Wang, F.J. Yu, High Expression of Serum miR-17-5p Associated with Poor Prognosis in Patients with Hepatocellular Carcinoma, Hepato-Gastroenterology 60(123) (2013) 549-552.
[73] J. Yu, K. Ohuchida, K. Mizumoto, H. Fujita, K. Nakata, M. Tanaka, MicroRNA miR-17-5p is overexpressed in pancreatic cancer, associated with a poor prognosis and involved in cancer cell proliferation and invasion, Cancer Biology & Therapy 10(8) (2010) 748-757.
[74] W. Li, J. Yi, X.J. Zheng, S.W. Liu, W.Q. Fu, L.W. Ren, L. Li, D.S.B. Hoon, J.H. Wang, G.H. Du, miR-29c plays a suppressive role in breast cancer by targeting the TIMP3/STAT1/FOXO1 pathway, Clinical Epigenetics 10 (2018).
[75] J.W. Rostas, H.C. Pruitt, B.J. Metge, A. Mitra, S.K. Bailey, S. Bae, K.P. Singh, D.J. Devine, D.L. Dyess, W.O. Richards, J.A. Tucker, L.A. Shevde, R.S. Samant, microRNA-29 negatively regulates EMT regulator N-myc interactor in breast cancer, Molecular Cancer 13 (2014).

[76] L. Wang, W. Gao, F. Hu, Z.Y. Xu, F.Q. Wang, MicroRNA-874 inhibits cell proliferation and induces apoptosis in human breast cancer by targeting CDK9, Febs Letters 588(24) (2014) 4527-4535.
[77] F. Xing, S. Sharma, Y. Liu, Y.Y. Mo, K. Wu, Y.Y. Zhang, R. Pochampally, L.A. Martinez, H.W. Lo, K. Watabe, nmiR-509 suppresses brain metastasis of breast cancer cells by modulating RhoC and TNF-alpha, Oncogene 34(37) (2015) 4890-4900.
[78] P. Du, X.P. Luan, Y.W. Liao, Y.T. Mu, Y. Yuan, J.X. Xu, J.J. Zhang, MicroRNA-509-3p inhibits cell proliferation and invasion via downregulation of X-linked inhibitor of apoptosis in glioma, Oncology Letters 15(1) (2018) 1307-1312.
[79] W. Dai, J. He, L. Zheng, M. Bi, F. Hu, M. Chen, H. Niu, J. Yang, Y. Luo, W. Tang, M. Sheng, miR-148b-3p, miR-190b, and miR-429 Regulate Cell Progression and Act as Potential Biomarkers for Breast Cancer, J Breast Cancer 22(2) (2019) 219-236.