UNIVERSITY OF THESSALY

SCHOOL OF SCIENCES

DEPARTMENT OF COMPUTER SCIENCE AND BIOMEDICAL INFORMATICS

Development of algorithms for the characterization of non-coding biomarkers in pathogenesis

PhD DISSERTATION

Athanasios Alexiou

Supervisor: Artemis G. Hatzigeorgiou,

Bioinformatics Professor

Lamia, 2023

DISSERTATION COMMITTEE

Artemis G. Hatzigeorgiou, Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly (Head of Doctoral Advisory Committee)

Pantelis Bagos, Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly (Member of Doctoral Advisory Committee)

Georgia Braliou, Assistant Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly (Member of Doctoral Advisory Committee)

Vassilis Plagianakos, Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly

Haralampos Karanikas, Assistant Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly

Panagiota Kontou, Assistant Professor, Department of Mathematics, University of Thessaly

Antonis Giannakakis, Assistant Professor, Department of Molecular Biology and Genetics, Democritus University of Thrace

Abstract

In the past fifteen years, advancements in technology, particularly Next-Generation Sequencing (NGS), have revolutionised biological and medical research by enabling cost-effective and large-scale biological data production. The integration of NGS datasets offers immense potential in understanding RNA biogenesis and functions, thereby enhancing research in areas such as identifying therapeutic targets and diagnostic biomarkers. The availability of vast amounts of publicly accessible data has bolstered data-driven research, with bioinformatics analysis playing a crucial role in characterising and comparing the abundance and interactions of both coding and non-coding RNA in various physiological and pathological states. MicroRNAs (miRNAs), which are small noncoding RNAs approximately 22 nucleotides long, are considered key regulators of gene expression at the post-transcriptional level. They are abundant in many organisms and play critical roles in a wide range of biological processes, both normal and disease-related. Over the past decade, extensive research has focused on investigating the role of miRNAs in complex diseases like cancer, with studies identifying specific miRNAs acting as either oncogenes or tumor suppressors. Small RNA sequencing (sRNA-Seq) has emerged as a powerful technique for quantifying these small noncoding RNAs on a large scale, providing valuable insights into the function of these important regulators.

This PhD dissertation is focused on the development of tools and resources allowing for the identification and study of noncoding biomarkers in various disease states. More specifically, the DIANA-microRNA-Analysis-Pipeline (DIANA-mAP), a fully automated computational pipeline was developed to facilitate the analysis of miRNA Next-Generation Sequencing (NGS) data. DIANA-mAP processes raw sRNA-Seq libraries aiming to perform quantification and Differential Expression Analysis. The pipeline places particular emphasis on the crucial step of data pre-processing, which greatly impacts the reliability of the final results, and even allows for automatic adapter inference. In an era with vast publicly available data without the necessary metadata information required for proper analysis, the latter highly increases opportunities for public data integration in studies. Through comprehensive evaluation, it demonstrates robust results and even superior performance compared to similar tools under adapter-agnostic analysis scenarios. Similarly, the DIANA-RSeq is an automated pipeline for the quantification of RNA sequencing data. It is composed of multiple standalone multi-option

modules that provide a number of possible analysis workflows utilising the state-of-the-art bioinformatic tools.

The resources created in this dissertation aimed to systematically collect and curate vast amounts of information from numerous studies in a precise and consistent manner. These resources provide various aspects in the context of noncoding biomarkers, including human tissue specific miRNA expression data, experimentally validated biomarker functionality, computationally predicted miRNA-RNA interactions, and even associations between microbes and diseases. By offering a user-friendly interface for querying the comprehensive results of numerous studies, these resources enable researchers to obtain structured and interconnected information. This valuable capability is expected to greatly assist in addressing intricate biological inquiries, formulating research hypotheses, and advancing scientific investigations.

During the course of my PhD studies I participated in 5 published studies (i.e. one on the development of a miRNA analysis pipeline tool, two on the development of resources providing curated published data for experimentally supported circulating miRNA biomarkers and bacterial-disease interactions, two on the development of resources providing miRNA expression data on human tissues and computationally predicted miRNA-RNA targets). I presented research results in 5 scientific conferences (3 international and 2 national) and contributed to the organisation of the European Conference of Computational Biology (ECCB 2018).

Table of Contents

1. Introduction

1.1. RNA Biology

RNA biology is a field of study that focuses on understanding the diverse roles and functions of RNA molecules within cells. RNA (ribonucleic acid) is a versatile molecule that plays crucial roles in gene expression, cellular processes, and regulation. It is involved in various aspects of cell biology, including transcription, translation, RNA processing, RNA modification, and RNA-protein interactions.

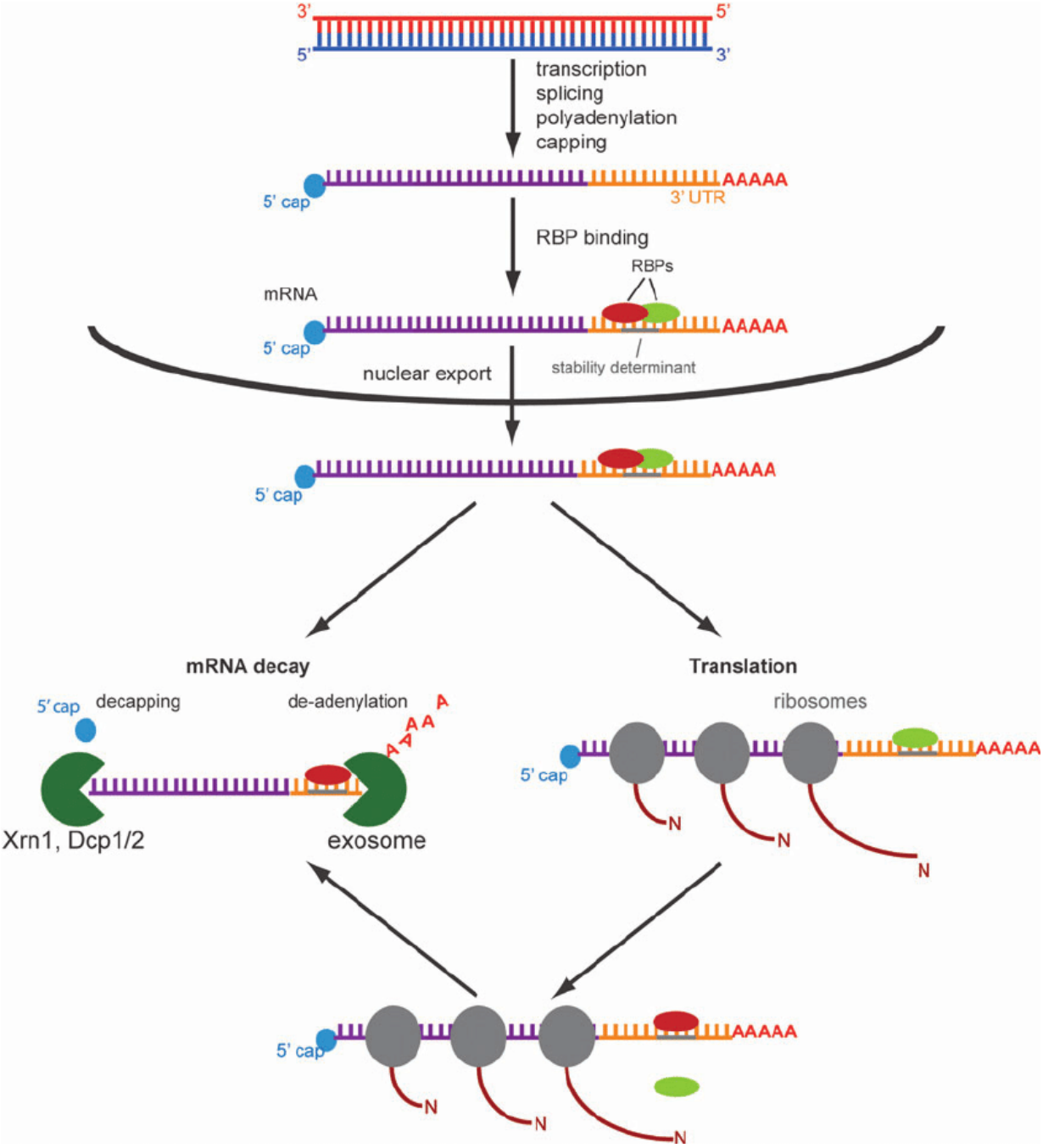

Figure 1. *General schematic representation of the life cycle of an mRNA* [1]

RNA molecules undergo several processing steps to become functional. These processes include capping, splicing, and polyadenylation in some cases. Additionally, they can undergo various modifications, such as methylation, pseudouridylation, and base modifications, which can impact their stability, localization, and functions. Apart from their standalone procession, RNA molecules interact with proteins to form ribonucleoprotein complexes (RNPs) that participate in various cellular processes. RNA-protein interactions are crucial for RNA stability, transport, translation, and regulation. Examples include ribosomes, spliceosomes, and RNA-binding proteins involved in RNA metabolism and regulation. [2] Regulation is also an important use case for RNA molecules, where they play a central role in gene regulation by controlling the expression of genes at the transcriptional and post-transcriptional levels [3]. They can influence gene expression through mechanisms such as transcriptional interference, chromatin remodelling, alternative splicing, RNA decay, and translation regulation. The RNA molecules are separated into different types usually based on their function and/or characteristics such as length, structure etc (see 1.1.1).

Moreover, and based on the above, advances in RNA biology have led to the development of RNA-based therapeutics, such as antisense oligonucleotides, small interfering RNAs (siRNAs), and RNA-based vaccines. These approaches utilise the specificity of RNA molecules to target and modulate gene expression in various diseases, including cancer, genetic disorders, and viral infections. [4][5]

Overall, RNA biology is a rapidly evolving field with extensive research focused on understanding the intricate functions and regulatory mechanisms of RNA molecules. It plays a fundamental role in deciphering the complexity of cellular processes and has significant implications for various areas of biology and medicine.

1.1.1. RNA types

There are multiple ways to categorise the RNA molecules such as their potential to directly code for proteins. However, RNA molecules are most usually classified into different types based on their functions and characteristics. The major classes of RNA include:

- **Messenger RNA (mRNA)**:
  mRNA acts as an intermediate messenger between DNA, which contains the genetic code, and the protein synthesis machinery. It carries the instructions encoded in DNA's nucleotide sequence in the form of codons, each of which specifies a particular amino acid. During translation, ribosomes read the codons on the mRNA and assemble the corresponding amino acids into a polypeptide chain, forming a protein. [6]

- **Transfer RNA (tRNA)**:
  Each tRNA molecule is specific to a particular amino acid and has a three-nucleotide sequence called the anticodon that is complementary to the corresponding codon on mRNA. It serves as a molecular adapter in the protein synthesis process, delivering specific amino acids to the ribosomes based on the information encoded in mRNA. Its accurate recognition and placement of amino acids are essential for the proper synthesis of functional proteins. [7]

- **Ribosomal RNA (rRNA)**:
  rRNAs form the structural and functional core of ribosomes, the cellular machinery responsible for translating mRNA into proteins. They act as a scaffold, facilitating the accurate assembly of amino acids into a polypeptide chain. rRNA molecules are highly conserved across different organisms and play a critical role in maintaining the structure and function of ribosomes. [8]

- **Small Nuclear RNA (snRNA)**:
  snRNAs are small RNA molecules, found in the nucleus of eukaryotic cells, that form part of snRNP complexes and play a critical role in the splicing of pre-messenger RNA (pre-mRNA) during the process of gene expression. They ensure the accurate removal of introns and the proper joining of exons to generate mature mRNA molecules. [9]

- **MicroRNA (miRNA)**:

  miRNAs are small non-coding RNAs that are transcribed in the nucleus of the cell and undergo their maturation process in the cytoplasm. Their main role is the regulation of gene expression by targeting specific mRNA molecules for degradation or translational repression. miRNAs are highly conserved amongst species which further validates their robustness as post-transcriptional regulators. [10]

- **Long Non-coding RNA (lncRNA)**:

  A diverse class of non-coding RNAs more than 200 nucleotides in length. They have been found to participate in various regulatory processes on multiple levels, including chromatin organisation and epigenetic regulation, transcriptional and post-transcriptional regulation as well as cellular signaling. The roles of the diverse subcategories of the lncRNA are an active research field and they are already known to interact with both mRNAs as well as miRNAs. [11][12]

1.1.2. miRNA

MicroRNAs (miRNAs) are small, non-coding RNA molecules that play a crucial role in post-transcriptional gene regulation. They are approximately 20-24 nucleotides in length and are found in plants, animals, and even some viruses [13]. miRNAs have been implicated in various biological processes, including development, cell proliferation, homeostasis, and disease. [3]

The biogenesis of miRNAs begins with their transcription by RNA polymerase II or III, producing primary miRNA transcripts (pri-miRNAs). These pri-miRNAs undergo processing in the nucleus by a complex called the microprocessor, which consists of the RNAse III enzyme Drosha and its cofactor DGCR8. The microprocessor cleaves the pri-miRNA to generate precursor miRNAs (pre-miRNAs), hairpin-like structures with a stem-loop configuration. [10][14]

Pre-miRNAs are exported from the nucleus to the cytoplasm, where they are further processed by an RNAse III enzyme called Dicer. Dicer cleaves the pre-miRNA, generating a double-stranded RNA duplex. One strand of the duplex, known as the guide strand or mature miRNA, is loaded into the RNA-induced silencing complex (RISC), which consists of Argonaute proteins (AGO). The guide strand of the miRNA guides RISC to target mRNAs through sequence complementarity. It has been reported that more than half mammalian mRNAs are miRNA targets [15]. [10][14]

Once bound to target mRNAs, miRNAs primarily exert their regulatory effects through two mechanisms: mRNA degradation and translational repression. If the miRNA is almost perfectly complementary to the target mRNA, it can lead to mRNA degradation through deadenylation and exonucleolytic decay. In cases where the miRNA has partial complementarity to the target mRNA, it typically results in translational repression, preventing the translation of the mRNA into a protein. [10]

miRNAs are known to target multiple mRNAs, and a single mRNA can be targeted by multiple miRNAs, allowing for complex regulatory networks and cross-talk between different pathways. The specificity of miRNA targeting is determined by the complementarity between the miRNA and its target mRNA, primarily involving recognition of the 3' untranslated region (UTR) of the mRNA.

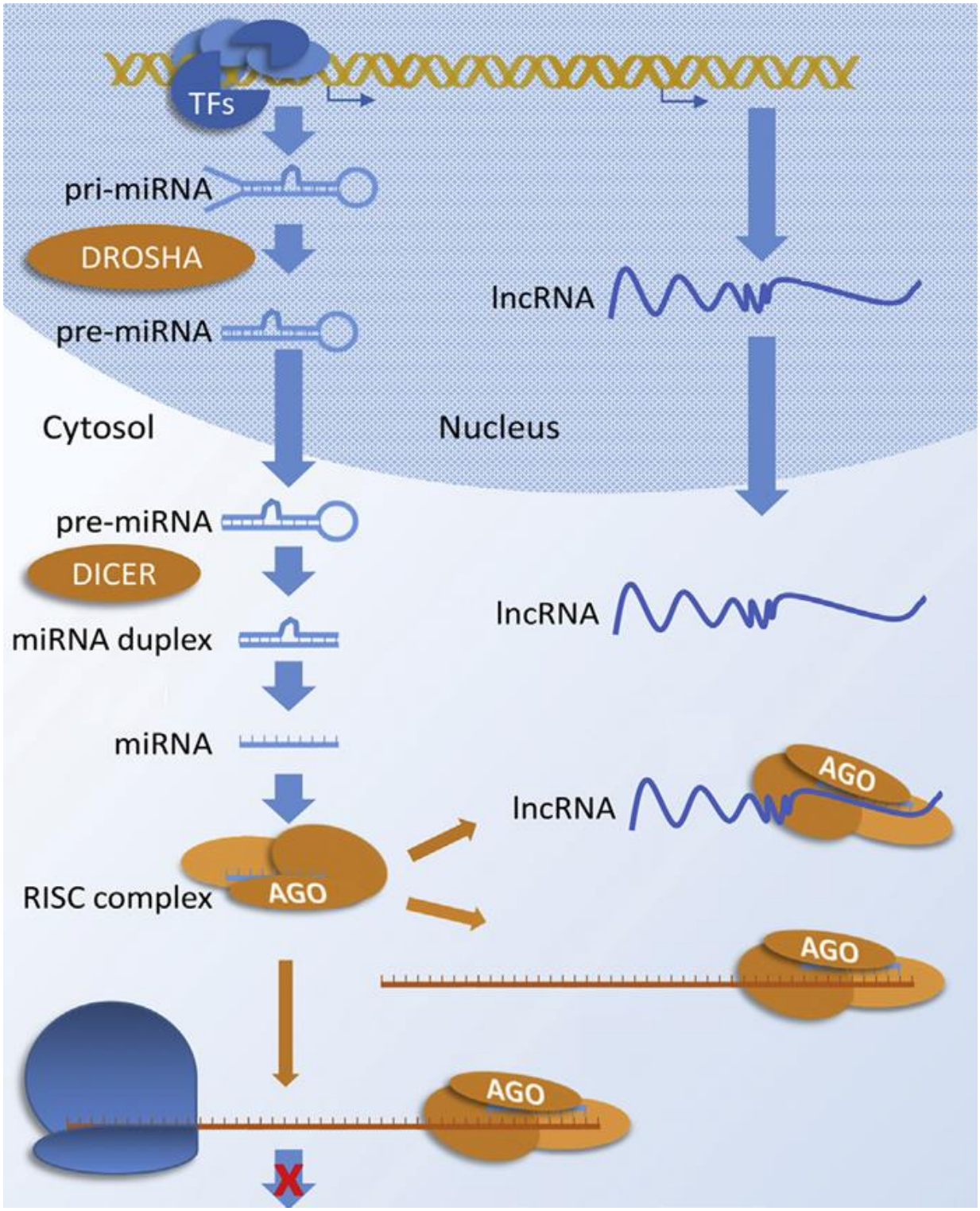

Figure 2. *Schematic representation of miRNA biogenesis and function* [16]

miRNAs have been implicated in various biological and disease processes, including cancer, cardiovascular diseases, neurodegenerative disorders, and immune responses [17][18]. Dysregulation of miRNA expression or function can lead to aberrant gene expression patterns and contribute to disease development and progression [19]. Moreover, they are often bound to

proteins and can be packaged in vesicles, which helps protect them from degradation [20]. Due to that, they are valuable as biomarkers for prognosis, diagnosis or monitoring for various diseases but also potential therapeutic targets for drugs, provided they carry discriminatory capacity [4].

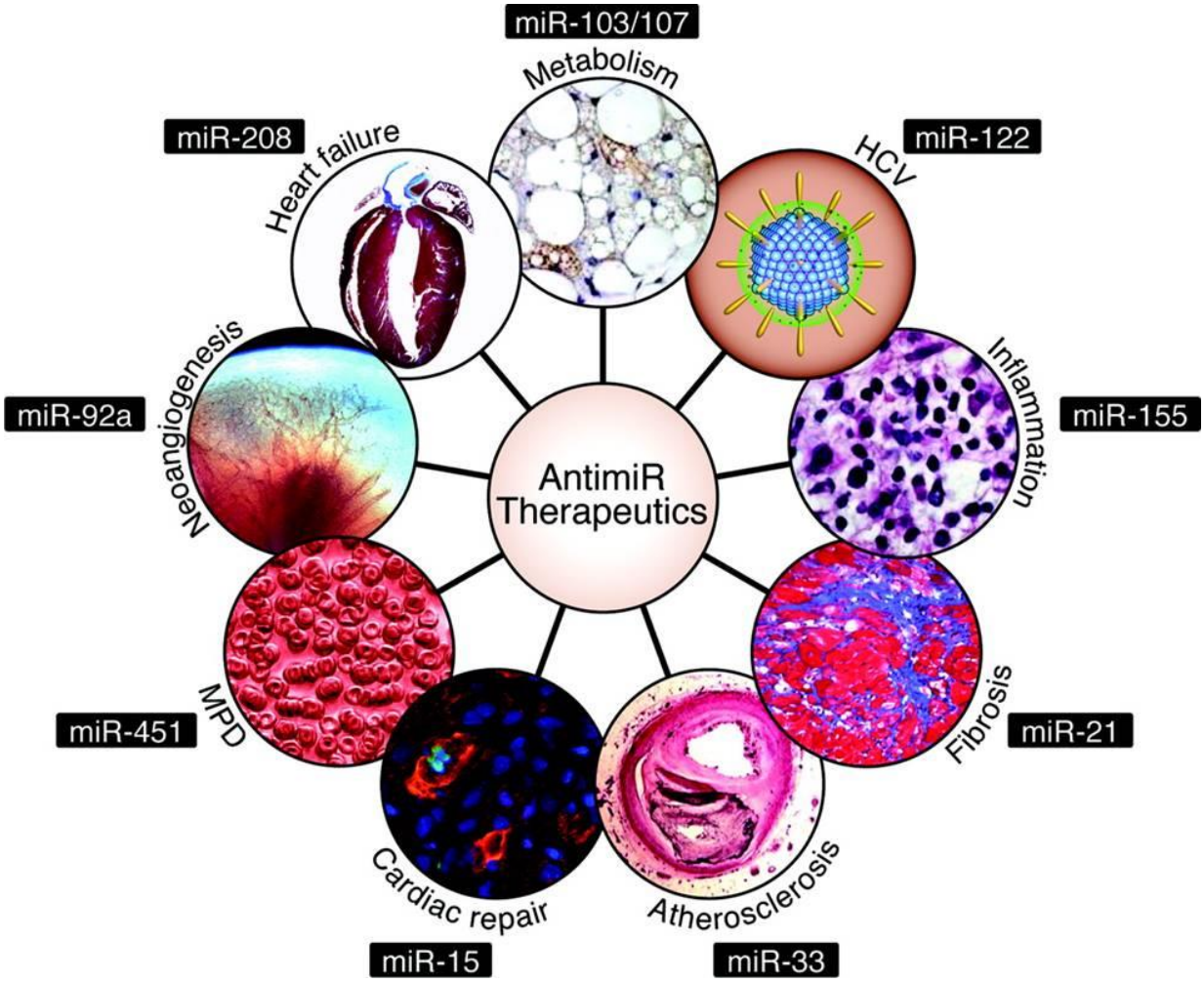

Figure 3. *Known miRNAs being pursued as clinical candidates for therapeutic targets in drugs* [4]

1.2. Biomarkers

Biomarkers are measurable indicators or characteristics that can be used to assess or evaluate biological processes, disease states, or responses to treatment. They can be molecules, such as proteins, nucleic acids, or metabolites, as well as physical characteristics, such as blood pressure or heart rate. Biomarkers are typically detected and measured through various techniques, including laboratory tests, imaging scans, or clinical assessments and can be found in various biological materials, including blood, urine, tissue, and genetic material.

Biomarkers can be categorised into different types based on their nature and function:

- **Genetic Biomarkers**: These biomarkers involve specific genetic variations or mutations associated with certain diseases or drug responses. Examples include BRCA1 and BRCA2 mutations linked to breast and ovarian cancer susceptibility [21].
- **Protein Biomarkers**: These biomarkers involve the measurement of specific proteins or peptides in biological samples. For example, prostate-specific antigen (PSA) is a protein biomarker used for prostate cancer screening and monitoring [22].
- **Metabolic Biomarkers**: These biomarkers reflect alterations in metabolic pathways and can be measured in body fluids. Examples include blood glucose levels used for diabetes diagnosis and monitoring.
- **Imaging Biomarkers**: These biomarkers involve the use of medical imaging techniques to visualise and quantify anatomical or physiological changes associated with diseases. Examples include brain amyloid imaging in Alzheimer's disease or tumor size measurements in cancer [23].
- **Cellular Biomarkers**: These biomarkers assess changes in specific cell populations or characteristics. For instance, CD4 cell count is a cellular biomarker used to monitor the progression of HIV infection [24].

These are just a few examples, and biomarkers play a crucial role in medicine and biomedical research. They can be used for various purposes, including:

- **Diagnosis**: Biomarkers can help identify the presence or absence of a particular disease or condition. For example, specific proteins in the blood can serve as biomarkers for certain types of cancer.

- **Prognosis**: Biomarkers can provide information about the likely course or outcome of a disease. They can help predict disease progression, response to treatment, or potential complications.
- **Treatment selection**: Certain biomarkers can assist in determining the most appropriate treatment approach for a particular individual. Genetic biomarkers, for instance, can help guide the choice of medications or therapies that are most likely to be effective and safe for a patient.
- **Monitoring**: Biomarkers can be used to track disease progression or treatment response over time. They allow healthcare professionals to assess the effectiveness of interventions and make adjustments if necessary.
- **Research and drug development**: Biomarkers are utilised in clinical trials and research studies to evaluate the safety and efficacy of new drugs, therapies, or interventions. They can help identify suitable candidates for participation and measure treatment effects.

Biomarkers are an active area of research, and new biomarkers are continually being discovered and validated. They are valuable tools for understanding and managing diseases, as well as guiding therapeutic interventions and personalised medicine approaches.

1.3. Next-generation Sequencing

The digitization of genetic material has revolutionised the analysis of biological data, ushering a new era of understanding in the field of biology and life [25]. Sequencing techniques brought this digitization but it was the arrival of fast, high-throughput and low-cost techniques that really opened the door to this new era through the possibility of analysis in scales unimaginable before [26].

Next-generation sequencing (NGS), also known as high-throughput sequencing, refers to a set of technologies that enable rapid and cost-effective sequencing of DNA or RNA. It represents a significant advancement over traditional Sanger sequencing [25], allowing researchers to analyse large amounts of genetic material in a shorter time frame and at a lower cost. [27]

NGS technologies have revolutionised genomics research, enabling a wide range of applications in fields such as biomedical research, clinical diagnostics, personalised medicine, evolutionary biology, and agriculture [28][29][30]. Below are the key steps involved in every next-generation sequencing process:

1. **Extraction and Sample Preparation**: Initially, the genetic material, which can be DNA or RNA, is extracted from the sample of interest. The DNA or RNA is then fragmented into smaller pieces, which can range from a few hundred to several hundred base pairs in length.
2. **Library Preparation**: The fragmented DNA or RNA is prepared into a sequencing library. This involves attaching specific DNA adapters or barcodes to the ends of the fragments. These adapters serve as recognition sites for sequencing platforms and enable multiplexing, where multiple samples can be sequenced in parallel in a single sequencing run in order to fully utilise the sequencing cycle and lower costs.
3. **Sequencing**: The prepared library is loaded onto a sequencing platform, belonging to one of several companies such as Illumina, Ion Torrent, or Pacific Biosciences. Each platform employs different methodologies but generally follows a similar principle: the DNA or RNA fragments are immobilised, amplified, and then sequenced simultaneously. Fluorescently labelled nucleotides or other detection methods are used to determine the sequence of bases in the DNA or RNA fragments.
4. **Data Analysis**: Following sequencing, the resulting data consists of millions or billions of short DNA or RNA sequences, known as reads. These reads need to be aligned to a

reference genome or assembled *de novo* to reconstruct the original genetic information. Bioinformatics tools and computational analysis are employed to process, interpret, and derive meaningful insights from the data. [31]

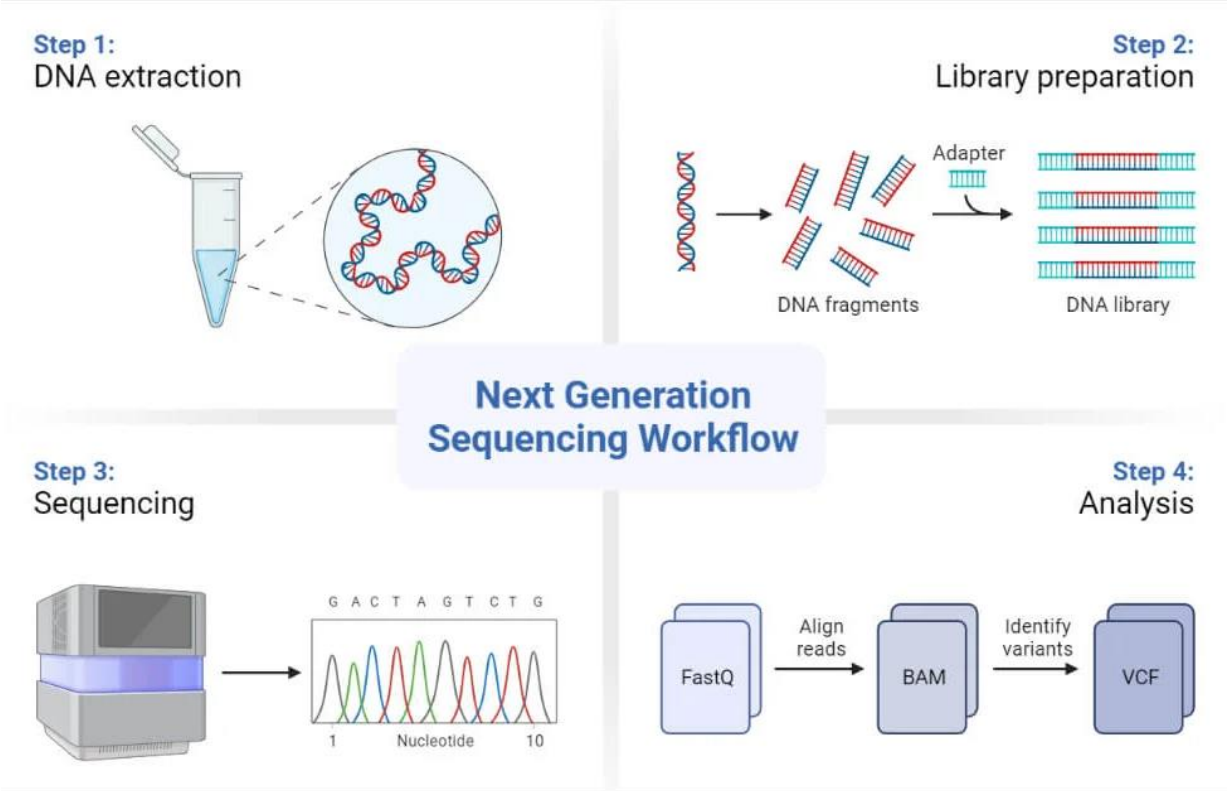

Figure 4. *Next-generation Sequencing steps* [32]

Next-generation sequencing technologies offer several advantages over traditional sequencing methods. They can generate massive amounts of data, providing higher throughput and enabling the sequencing of entire genomes, exomes (protein-coding regions of the genome), or transcriptomes (all RNA molecules) [29]. NGS also allows for the detection of genetic variations, including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variations. It has contributed significantly to our understanding of the genetic basis of diseases, the identification of novel genetic markers, and the discovery of new drug targets [28].

As technology continues to advance, new generations of sequencing platforms are being developed, offering improved accuracy, longer read lengths, and decreased costs. These advancements are expanding the scope of NGS applications and facilitating the translation of genomics into clinical practice. [33]

1.3.1. RNA-Sequencing

RNA-Seq, or RNA sequencing, is a high-throughput sequencing technique used to study and analyse the transcriptome of an organism. The transcriptome refers to the complete set of RNA molecules, including messenger RNA (mRNA), non-coding RNA, and small RNA molecules, produced in a particular cell or population of cells. [34]

RNA-Seq involves the conversion of RNA molecules into complementary DNA (cDNA), followed by sequencing of the cDNA fragments. The process begins with the extraction of total RNA from the cells or tissues of interest. This RNA is then treated to remove ribosomal RNA (rRNA) and enrich for the messenger RNA (mRNA) fraction, which represents the protein-coding genes and is the primary focus of analysis in most RNA-Seq experiments. Next, the mRNA molecules are converted into cDNA using reverse transcription. This cDNA synthesis step incorporates specific sequences called adapters, which allow for subsequent sequencing and identification of individual cDNA molecules. [34]
Once the cDNA library is prepared, it is subjected to high-throughput sequencing using platforms such as Illumina, Ion Torrent, or Pacific Biosciences. During sequencing, the cDNA molecules are fragmented, and the sequences of the fragments are determined in a massively parallel manner. The resulting sequence reads are typically short, ranging from a few dozen to a few hundred nucleotides in length. After sequencing, the raw data undergoes a series of bioinformatics analyses to process, map, and quantify the reads in order to assess gene expression. [34][35]

RNA-Seq has revolutionised transcriptomic research as it provides a comprehensive and quantitative view of gene expression. It allows researchers to study gene expression levels, identify novel transcripts, discover alternative splicing events, detect non-coding RNA molecules, and investigate gene fusion events. RNA-Seq has numerous applications in biological and biomedical research, including the study of development, disease mechanisms, drug discovery, and personalised medicine. [35]

1.3.2. small RNA-Sequencing

Small RNA-Sequencing, or Small RNA-Seq, is a specialised RNA sequencing technique used to study and analyse small RNA molecules in a sample. Small RNAs are a class of regulatory

RNA molecules that are typically less than 200 nucleotides in length. They play crucial roles in gene expression regulation, post-transcriptional gene silencing, and other cellular processes.

Small RNA-Seq follows a similar general workflow as conventional RNA-Seq, but with specific adaptations to capture and sequence small RNA molecules. The most important differences are highlighted below:

1. **Target RNA Molecules**: RNA-Seq focuses on the analysis of the entire transcriptome, including both protein-coding and non-coding RNA molecules. It captures and sequences a wide range of RNA species, such as messenger RNA (mRNA), long non-coding RNA (lncRNA), and other non-coding RNAs. In contrast, Small RNA-Seq specifically targets and analyses small RNA molecules, which are typically less than 200 nucleotides in length. This includes small regulatory RNAs like microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs). [36]
2. **Library Preparation**: In RNA-Seq, the library preparation protocol involves enrichment for the mRNA fraction, which represents the protein-coding genes. This involves removing ribosomal RNA (rRNA) and other non-mRNA molecules to enhance the coverage of mRNA transcripts. The cDNA library prepared for RNA-Seq is generally generated from poly(A)-enriched RNA. On the other hand, Small RNA-Seq library preparation involves specific size selection to enrich small RNA molecules, and adapters are ligated directly to the small RNAs without the need for poly(A) selection. [35][36]
3. **Sequencing Depth and Read Length**: Due to the broader scope of the transcriptome, RNA-Seq typically requires higher sequencing depth and longer read lengths compared to Small RNA-Seq. RNA-Seq often aims for higher coverage to accurately quantify gene expression levels, detect low-abundance transcripts, and analyse alternative splicing events. In contrast, Small RNA-Seq can achieve sufficient coverage and resolution with lower sequencing depth and shorter read lengths since small RNA molecules are typically more abundant and do not require full-length sequencing. [35][36][37]

1.4. Bioinformatic Analysis

1.4.1. Biological Data Analysis Workflow

The advances in the technological sector the past 40 years has revolutionised the field of biology in the acquisition and accumulation of data. That amount of data made the manual analysis largely obsolete due to scale and the involvement of the field of Informatics for the biological data analysis mostly gave birth to the Bioinformatics field as a whole. A typical biology experiment nowadays begins in a laboratory with the extraction of the sample(s), the library preparation and their sequencing. From the moment the sequencing platform digitises the biological information, the data generated signify the start of the bioinformatic involvement in the experiment. The bioinformatic analysis mostly consists of preparing the data for the analysis and then analysing them in accordance with the genomic data of the reference organism in study when possible, or a *de novo* reference is created based on the data itself. The data is usually analysed from multiple angles trying to identify genomic differences between samples in different time points or conditions depending on the input as well as the experiment objectives. Finally, the results are to be interpreted for their biological or clinical meaning, validating or contradicting the original hypothesis and leading to possibly re-analysing the data in different ways, based on the observations, or the design of new experiments.

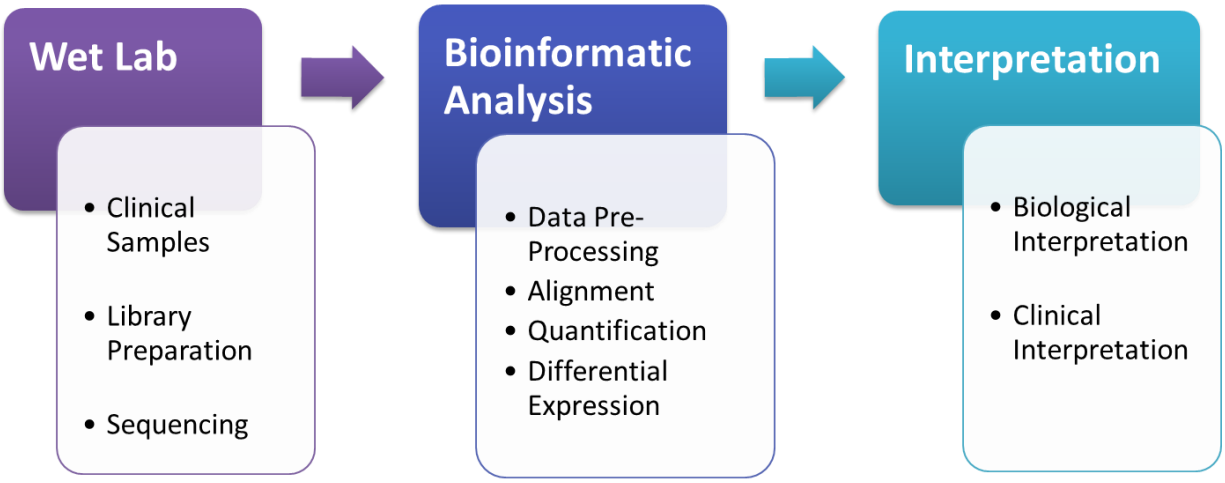

Figure 5. *Biological data analysis workflow steps (figure created for the needs of the current thesis)*

1.4.2. Typical Bioinformatic Analysis

A typical bioinformatic analysis workflow is usually composed of four main steps. There are many bioinformatic tools that implement the various steps or whole pipelines that perform the whole analysis.

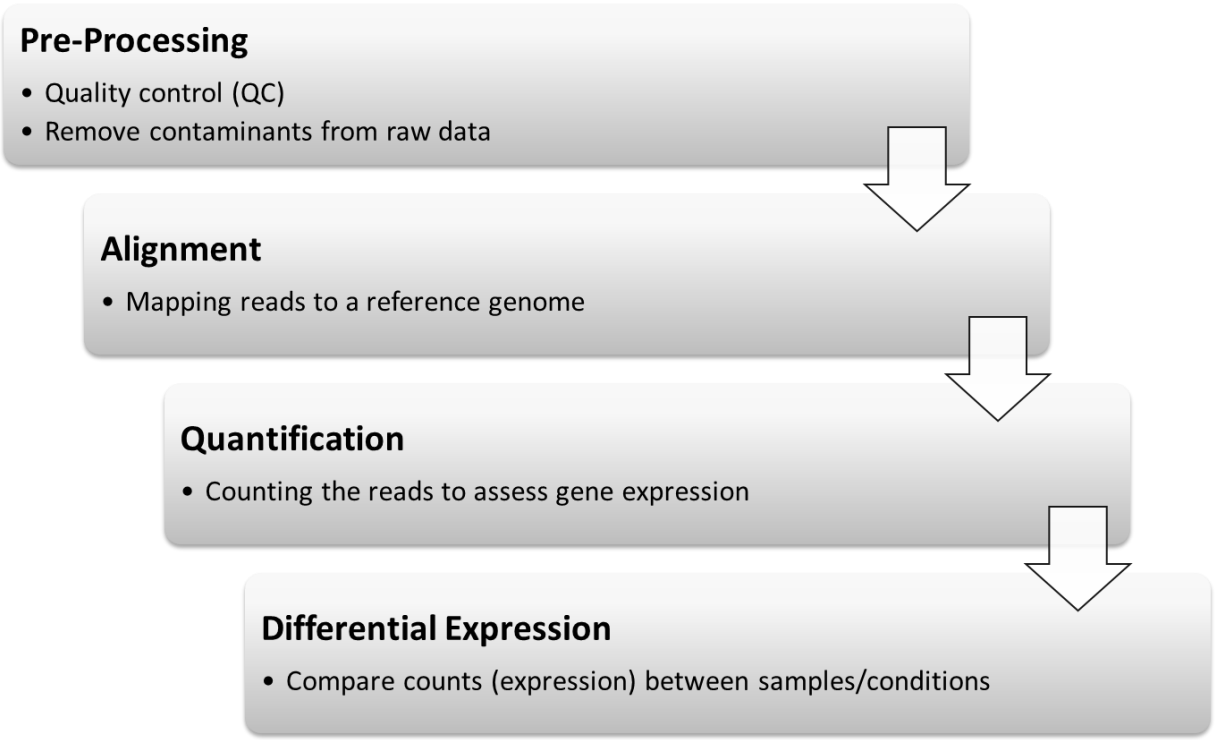

Figure 6. *Typical bioinformatic analysis workflow steps (figure created for the needs of the current thesis)*

The raw data input is provided by the sequencing platform and first this data is pre-processed. This step aims to assess both input and output data quality, remove low quality reads or nucleotides at the end of them while also removing adapters, other contaminants and very short reads. The alignment step that follows, uses the reference genome of the organism in study in order to align the input reads to it as an extra quality measurement. That way we have a better understanding of which of the data actually belong to the organism of study and which might be contaminants or other organisms (eg. bacteria, viruses etc.). Using only the successfully aligned reads, the quantification step aims to count the reads based on their alignment point. For this, an annotation of the reference organism is needed and the result is a table with the genes of the organism in study and a count of reads front the data aligned on each one. Finally, one of the most common analysis steps after the quantification process is the differential expression. Using the counts created in the previous step for multiple samples of different

conditions or time points, the genomic differences between those states is measured by calculating the differences between those counts, in sample groups, through statistical methods.

1.4.3. Small RNA-Sequencing Analysis Tools

In recent years, significant advancements in technology, particularly Next-Generation Sequencing (NGS), have resulted in the cost-effective production of large-scale data, transforming the fields of biology and medicine. This surge in sequencing data, including microRNAs (miRNAs), has made miRNA research a prominent area of investigation. NGS RNA sequencing has been widely adopted, with protocol pipelines like the "Tuxedo" pipeline [38] and the "new Tuxedo" package [39] serving as prevalent approaches for RNA-Seq expression analysis, offering a relatively straightforward methodology. The availability of extensive publicly accessible data has further bolstered data-driven research, highlighting the importance of efficient computational analysis for biological data. However, researchers who lack training or face challenges such as complex software installation or limited computational resources may miss out on utilising these valuable datasets and opportunities for research advancement. Moreover, the concise nature of miRNAs necessitates meticulous data preprocessing in sRNA-Seq analysis, as adapters can have a substantial impact on the reads due to their relative length compared to the reads. Additionally, the presence of numerous sequences that can map to multiple locations in the genome presents a challenge during the quantification step, which can potentially result in significant data misinterpretations. Over the years, numerous sRNA-Seq analysis pipelines have been developed, and they can be categorised based on their preprocessing approach to adapter identification. Many of these programs focus on specific analysis aspects, such as isomiR detection (sRNAnalyzer [40], QuickMIRSeq [41], Prost! [42], Jasmine [43], exogenous sequence and different non-coding RNA identification (sRNAnalyzer, sRNAtoolbox [44], mirTools 2.0 [45], or de novo miRNA identification (CAP-miRSeq [46], miRge 2.0 [47]). In terms of preprocessing, most of these tools assume the presence of a known adapter and perform basic adapter removal. However, when users don't have access to adapter information, this becomes a limitation. Some tools, like sRNAnalyzer, offer a selection of commonly used adapters as options when the adapter is unknown while also use sequence frequencies to infer primers in multiplexed datasets. MiARma-Seq [48] stands out as it provides comprehensive adapter-agnostic preprocessing

analysis by utilising Minion from the Kraken toolset [49] to infer the adapter using sequence frequencies. No further preprocessing steps are applied to eliminate potential remaining adapter contaminants from the data.

1.5. MicroRNA Resources

Building a thorough and cohesive human miRNA expression atlas presents challenges due to several factors. Firstly, the available datasets are provided in different formats, ranging from raw FASTQ files to count-level estimates and normalised expression values. Secondly, these datasets have been generated using different miRNA annotation sources and versions, leading to inconsistencies. Lastly, the quantification pipelines and algorithmic choices employed in the production of these datasets vary, further adding to the complexity. Additionally, the annotation of sample- and study-specific metadata in public repositories and by individual submitters lacks systematic standardisation, making it difficult to integrate datasets from diverse sources into a unified database. However, the task of cataloguing miRNA expression estimates, particularly in disease contexts, is a valuable scientific pursuit with various potential applications in both basic and translational research. Currently, there are several implementations available, each differing in their scope, coverage, and functionality. Examples include miRmine [50], HMED [51], DASHR2 [52], YM500v3 [53], SEAWeb [54], and DeepBase v3.0 [55]. These databases aim to capture a wide range of sample types and tissues, with varying numbers of datasets available.

miRmine, HMED, and DASHR2 encompass 304, 401, and 802 datasets, respectively. miRmine allows for single or multiple miRNA queries and provides sample-level RPM matrices for cell lines or tissues. DASHR2 focuses on the genomic localization of small RNAs and provides information on their sequence and secondary structure, along with a static expression table. YM500v3 specifically focuses on cancer, utilising samples solely derived from TCGA [56]. SEAWeb is a web application offering various analyses, such as differential expression and classification, enabling users to upload their own results for comparison against the database content. It includes 4258 datasets from 10 organisms, with a significant portion (3360) representing human datasets. DeepBase v3.0 categorises data into 'Cancer data' (from TCGA and ICGC [57]) and 'Tissue data' (500 publicly available datasets). However, it's important to note that miRNA expression values in DeepBase refer to precursors or miRNA host genes,

lacking abundance estimates of the functional mature miRNA forms in all analysis types provided. It is worth mentioning that HMED and YM500v3 were not operational at the time of writing. The existing databases mostly consist of a limited number of datasets or focus primarily on TCGA due to practical reasons. TCGA provides a relatively standardised resource, while studies from GEO[58]/SRA[59] exhibit extensive diversity in terms of library preparations, adapter usage, and sample quality.

There are also several databases available that contain information on circulating miRNAs in blood, serum, and plasma, with a focus on their abundance in various disease conditions. One such database is miRandola [60], which includes dysregulated extracellular miRNAs, long non-coding RNAs, and circular RNAs associated with diseases. Although miRandola provides manual annotations on biomarker validation methods, it lacks biomarker-specific query options and lacks metadata on statistical methods and cohorts, making it challenging to extract relevant information. The human miRNA-disease association (HMDD) database [61] collects comprehensive information on (epi-)genetics, target interactions, tissue expression, and circulatory associations of miRNAs with diseases. The circulatory category in HMDD focuses on the differential regulation of miRNAs in disease states and provides target-based functional enrichment analysis. Another database, the Circulating MicroRNA Expression Profiling (CMEP) database [62], utilises expression profiles from sequencing or microarray experiments on biofluids to offer online tools for analysing differential expression, pathway enrichment, and potential diagnostic marker identification.

2. Methods and Results

2.1. Tools

Bioinformatic tools are algorithms, pipelines or applications that are created to perform either parts of, or the complete analysis of biological data. They are geared toward specific data and are usually created to be treated as a black box by the end users.

2.1.1. DIANA-mAP

Overview

In an era where data is produced faster than it is analysed, more and more scientists use publicly available datasets for their analyses. Moreover, the low percentage of public data with the appropriate accompanied metadata leads to a higher need of adapter-agnostic pre-processing analysing tools. Mistakes in the pre-processing step of an analysis will undeniably produce erroneous results and possibly false conclusions. Due to the short length (~22nt) of miRNAs the impact of mishandlings in those early parts of their analysis is very high, yet to this day, very few tools provide more than simplistic removal of known adapters. DIANA-mAP [63] is a fully automated computational pipeline with a strong focus in the pre-processing step that allows for miRNA quantification and Differential Expression Analysis to be conducted in an easy, scalable, efficient, and intuitive way.

The tool performs reliable de novo pre-processing through an in-house algorithm utilising DNApi [64] to infer the sequencing adapters. It incorporates extra pre-processing loops to remove fragmented adapter contaminants while utilising a known-adapter-library mechanism allowing for user input on possible known adapters and cross reference with the inferred adapter results. All the above steps result in a more robust and efficient adapter identification and trimming process. Alignment to a reference genome is performed through Bowtie [65] and miRDeep2 [66] is utilised for the quantification process along with known miRNAs from miRBase [67] database. Finally, a differential expression analysis is optionally performed using DESeq2 [68]. The provided results at the end of the analysis include graphs and a detailed summary report containing all the important information concerning the analysis. The pipeline has been developed in R, utilising its rich supporting library sets, and can be run on POSIX systems (Mac, Linux, Unix, BSD). It is freely available through Github and Docker. My contribution to the project was on the development, testing and comparison of the pipeline while also being the first author of the publication.

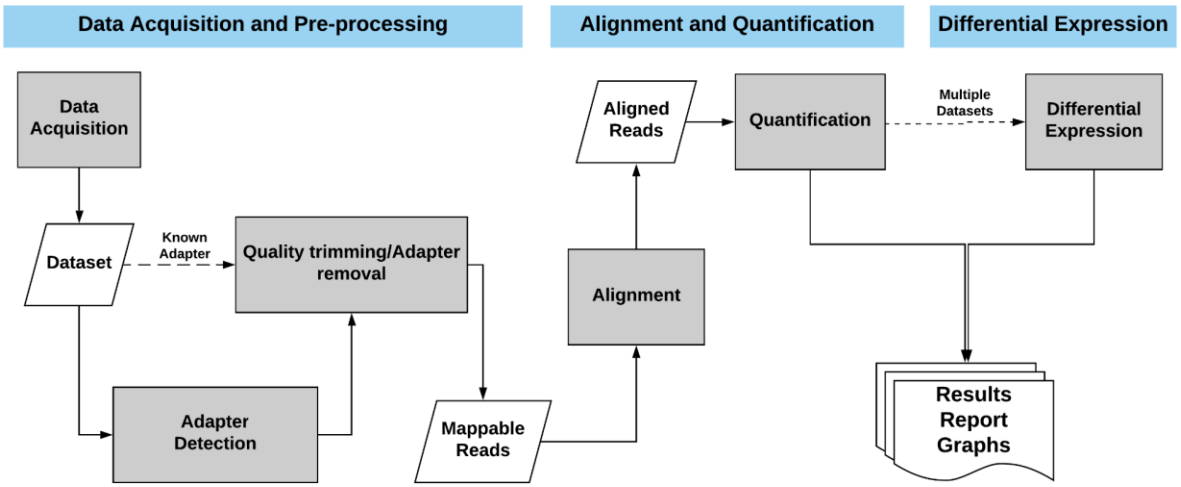

Figure 7. *The DIANA-microRNA-Analysis-Pipeline (DIANA-mAP) analysis workflow* [63]

The DIANA-mAP pipeline consists of three distinct modules as seen in Figure 7. Each module is composed of a number of steps that employ tools and algorithms aiming to perform the appropriate analysis on the imported data.

Data Acquisition and Pre-processing

The first step of the module is the *Data Acquisition*, an optional step allowing the user to select one of the three most common biological data repositories (SRA [59], GEO [58], ENCODE [69]) and download data by using the appropriate accession number. It allows for a faster streamlined data acquisition in the cases where the user also studies publicly available data apart from in-house ones.

The following step utilises FASTQC [70] to perform a first round of *Quality Control* in order to assess the general quality of the raw data imported. It provides a set of graphs and a report that provides the user the overall inspection and evaluation of the data before continuing with the computationally intensive parts of the analysis. In case of low quality, automated appropriate warnings and termination options are presented by the pipeline.

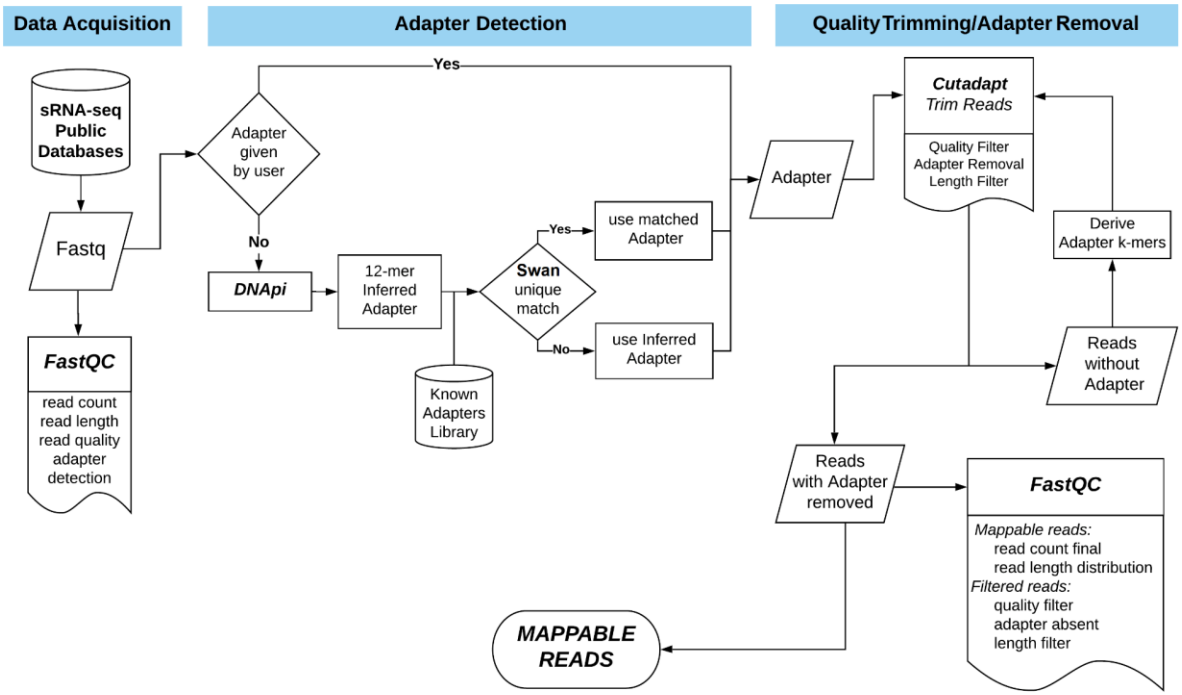

Figure 8. *The DIANA-mAP Pre-processing workflow* [63]

The third and most important step of the module is *Adapter Detection and Removal*. An in-house algorithm was developed for this step, which focuses on the detection and removal of adapter contaminants from the sequencing process. If the adapter sequence used during the sequencing is known to the user, it can be provided to the pipeline through configuration options and DIANA-mAP will continue to the removal part. Otherwise, the pipeline utilises DNApi [64] which uses the sequence frequency to infer the 3' adapter of the dataset. Additionally, a file containing a set of known adapter sequences is provided to the user which can be enriched by them. Through the use of Swan, a short sequence alignment tool from the Kraken software package [49], the DNApi inferred sequence is tested for similarity against the known adapter sequences and on a result of 90% or more the full known adapter sequence is used. If no match is found, DIANA-mAP uses the DNApi 12-mer inferred sequence as the adapter to be removed.

For the adapter removal part of the algorithm, Cutadapt [71] is used to remove the adapter sequence as well as low quality bases and even reads that are too long or short based on provided thresholds. This process however, only removes full instances of the adapter sequence. In order to tackle the fragmented adapter contaminants in the data, the adapter sequence is fragmented into k-mers (default: 10) and the above process is repeated for all of them. This loop process cleanses the data from lingering fragmented adapter contaminants,

which are usually leftover from traditional adapter removal practices, thus allowing more reads to be further processed downstream in the pipeline.

Finally, a second round of *Quality Control* is applied on the pre-processed data in order to assess the remaining data and the effect of the pre-processing on them before continuing with the analysis. Again, appropriate warnings are automatically generated by DIANA-mAP based on the quality statistics.

Alignment and Quantification

This module firstly initiates the *Alignment* step, where the pre-processed reads are aligned to the reference genome of the organism in study, as an extra quality assurance in order to identify which reads are clean and belong to the organism while also grouping the rest for other possible studies. For this step, the mapping script of miRDeep2 [66] is used, which invokes Bowtie [65] for its documented precision and focus on the alignment of smaller length reads. The *Quantification* step that follows, uses again miRDeep2 in order to align the reads provided by the previous step, but this time against the known miRNAs of the organism in study acquired from the miRBase DB [67]. This results in a count of input data reads to the known miRNAs of the organism, essentially providing an expression matrix for the known miRNAs.

Differential Expression

The final module of the pipeline is Differential Expression and is optionally executed when the pipeline has been employed for the analysis of multiple samples which represent different conditions or time points. In those cases, a table is provided assigning the sample data to the different conditions of study and DESeq2 [68] is utilised using the outputs of the quantification steps to calculate statistically significant differences in the expression of miRNAs on the different conditions.

Results

DIANA-mAP produces a number of results such as the expression table for the miRNAs, but also stores the intermediate and final data allowing for their use in different downstream analysis. A final report is created aiming to provide a quick overview of the analysis to the user but even more emphasis has been given to graphs. The length distribution graphs before and after the analysis (Figure 9A and B) provide a good look on how the data have shaped after the pre-processing, with the characteristic peak on the 22 bases for human data expected since most human miRNA are about 22 nucleotides long. A useful overview on the results of the first

module is also provided with the pie chart (Figure 9C) showing the distribution of the reads in data after pre-processing, where a high percentage of short reads or a low percentage of mappable reads could indicate an anomaly on the sample data with a glance. Finally, the differential expression module also produces results such as heat maps or even PCA graphs (Figure 9D) showing the clustering of samples based on their miRNA expression profile.

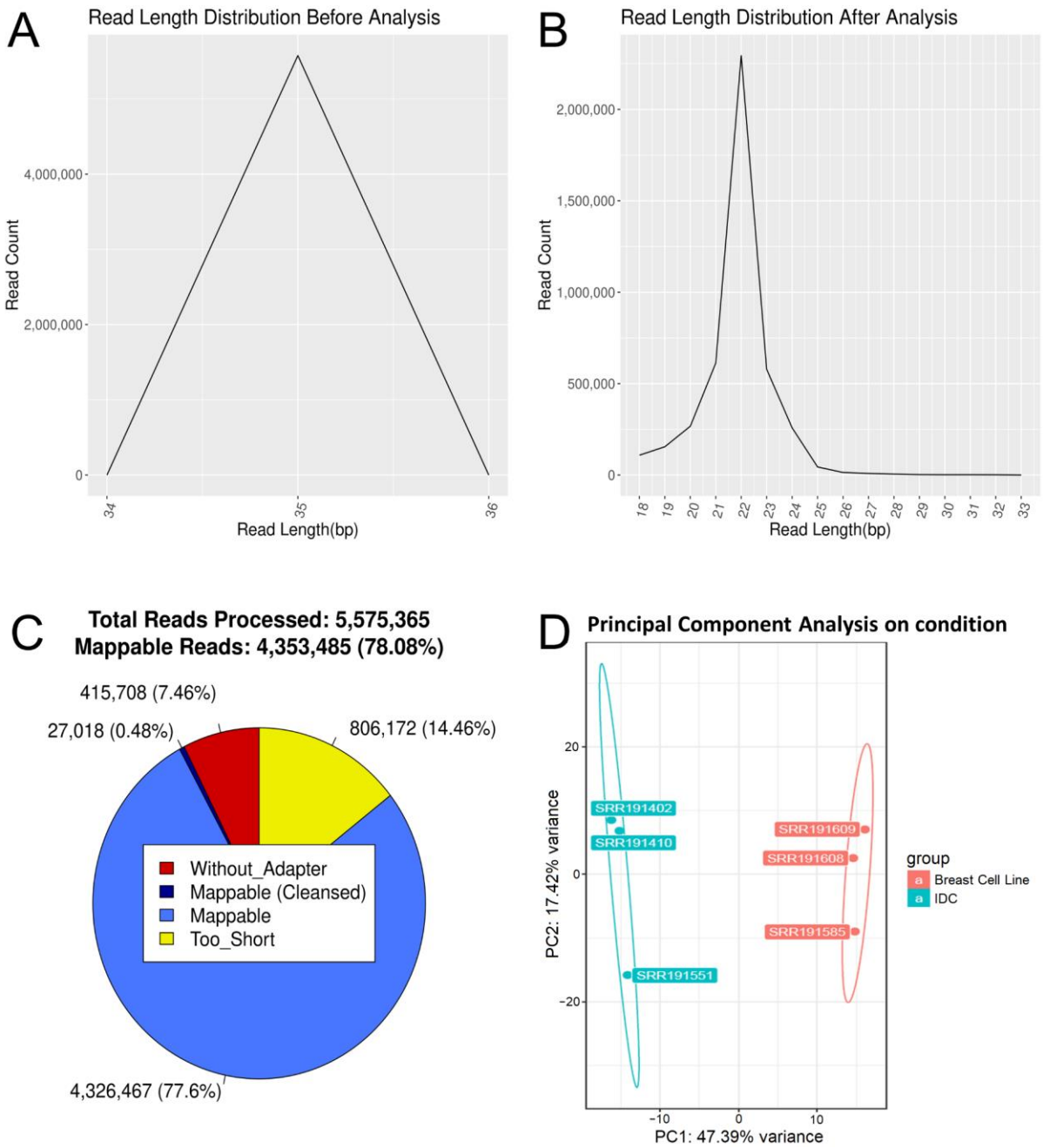

Figure 9. *DIANA-mAP visualisation results* [63]

Evaluation

In the past, preprocessing raw data was considered a simple step in data analysis, assuming knowledge of the adapters and/or primers used for sequencing. However, with the increasing speed of data production and the use of publicly available datasets, it has become common for information about the adapters/primer used in a dataset to be missing or incomplete. DIANA-mAP was developed with a strong emphasis on the preprocessing step, allowing analysis of datasets without prior knowledge of the adapter used. The way a sample is processed during preprocessing directly affects the quantity and quality of the sequencing reads, which subsequently impacts the entire analysis process. To evaluate the performance of DIANA-mAP in this regard, we compared its mapping, quantification, and time performance to miARma-Seq [48], the only miRNA analysis tool known to address this requirement.

For evaluation purposes, we used two dataset groups. The first group, named Dataset_Group_1, consists of eight sRNA-Seq libraries (GSE47602) from a study on the human MCF7 cell line that investigated miRNA regulation under hypoxia conditions [72]. Two files from this group were used for evaluation by miARma-Seq, while the other six were provided as examples upon download of the pipeline. The second group, named Dataset_Group_2, includes six sRNA-Seq datasets (GSE15229) originating from a study performed on normal and malignant human B cells [73], along with 18 additional datasets obtained from the Sequence Read Archive (SRA) and integrated into the functional transcriptomics tool DIANA-miExTra v2.0 [74]

Both pipelines, DIANA-mAP and miARma-Seq, were used without providing any adapter information for the data analysed. Additionally, none of the analysed dataset adapters were in the custom library with the known adapters provided by DIANA-mAP. The default parameter values were used for both tools, with a few exceptions: the trimming quality threshold was set to 20, the minimum and maximum allowed read lengths were set to 18 and 50, respectively, the maximum allowed mismatch on seed regions during genome alignment was set to 1, the reference genome used was GRCh37 (hg19), and the miRNA reference used was miRBase v20 [67]. The evaluation was performed on a single core of a High Performance Computer (HPC) with 48 cores operating at 2.3 GHz and 256 GB of RAM, running on the CentOS operating system. For comparison purposes, two metrics were chosen: the percentage of total reads mapped to the genome and the percentage of total reads that were mapped to known miRNAs, also referred to as quantified reads. These metrics provide essential information about the results of any analysis tool and are crucial for any further analyses such as Differential Expression.

In both groups of datasets, we identified six datasets (SRR873386, SRR873389 from Dataset_Group_1 and SRR033711, SRR033725, SRR033728, SRR033730 from Dataset_Group_2) where there was a notable disparity in the mapped and quantified reads between the two programs (Table 1). Across these six datasets, DIANA-mAP [63] quantified an average of 72.5% of the total reads, whereas miARma-Seq [48] only quantified an average of 1.3%. Upon closer examination of the results for these datasets, it became evident that miARma-Seq mishandled a significant portion of the reads during the preprocessing step. This was primarily due to false inference of adapters and/or excessively sensitive read cleansing, resulting in the discarding of too many reads that were deemed too short. For the remaining 26 datasets, both tools produced very similar results, with insignificant differences of less than 2% observed in the raw count numbers of quantified known miRNAs (Table 1).

Dataset Group 1							
Dataset Accession No.	**Number of Reads**	**Mapped Reads**			**Quantified Reads**		
		miARma-Seq	**DIANA-mAP**	**Difference (% of Total Reads)**	**miARma-Seq**	**DIANA-mAP**	**Difference (% of Total Reads)**
SRR873382	500000	410190 (82.04%)	395314 (79.06%)	**2.98**	384692 (76.94%)	377084 (75.42%)	**1.52**
SRR873383	500000	404071 (80.81%)	388416 (77.68%)	**3.13**	379041 (75.81%)	371648 (74.33%)	**1.48**
SRR873384	500000	376218 (75.24%)	368605 (73.72%)	**1.52**	363215 (72.64%)	358021 (71.60%)	**1.04**
SRR873385	500000	365339 (73.07%)	359338 (71.87%)	**1.2**	346597 (69.32%)	345055 (69.01%)	**0.31**
SRR873386	500000	10467 (2.09%)	406743 (81.35%)	**79.26**	7915 (1.58%)	391566 (78.31%)	**76.73**
SRR873387	500000	416391 (83.28%)	406197 (81.24%)	**2.04**	389770 (77.95%)	389028 (77.81%)	**0.15**
SRR873388	500000	408720 (81.74%)	400364 (80.07%)	**1.67**	389295 (77.86%)	386969 (77.39%)	**0.47**
SRR873389	500000	3182 (0.64%)	415122 (83.02%)	**82.39**	2529 (0.51%)	399397 (79.88%)	**79.37**

Table 1. *Comparison of raw miRNA read results for DIANA-mAP and miARma-Seq on Dataset Group 1 without adapter information [63]*

Out of the total of 32 samples, information regarding the adapters used in the original study was available for 18 samples. DIANA-mAP accurately predicted all those 18 adapters using DNApi's exhaustive mode, demonstrating a high level of precision in automatic adapter detection. In the remaining 14 samples, a high percentage of adapter detection (>87%) and high mapping and quantification percentages strongly indicate that DIANA-mAP successfully inferred the adapters (Table 1). On the other hand, miARma-Seq displayed lower trimming percentages and discrepancies in predicted adapters for 6 out of the 32 samples, providing insights into the performance differences observed. Figure 4 illustrates the quantification results for both groups of datasets in the form of a scatter plot.

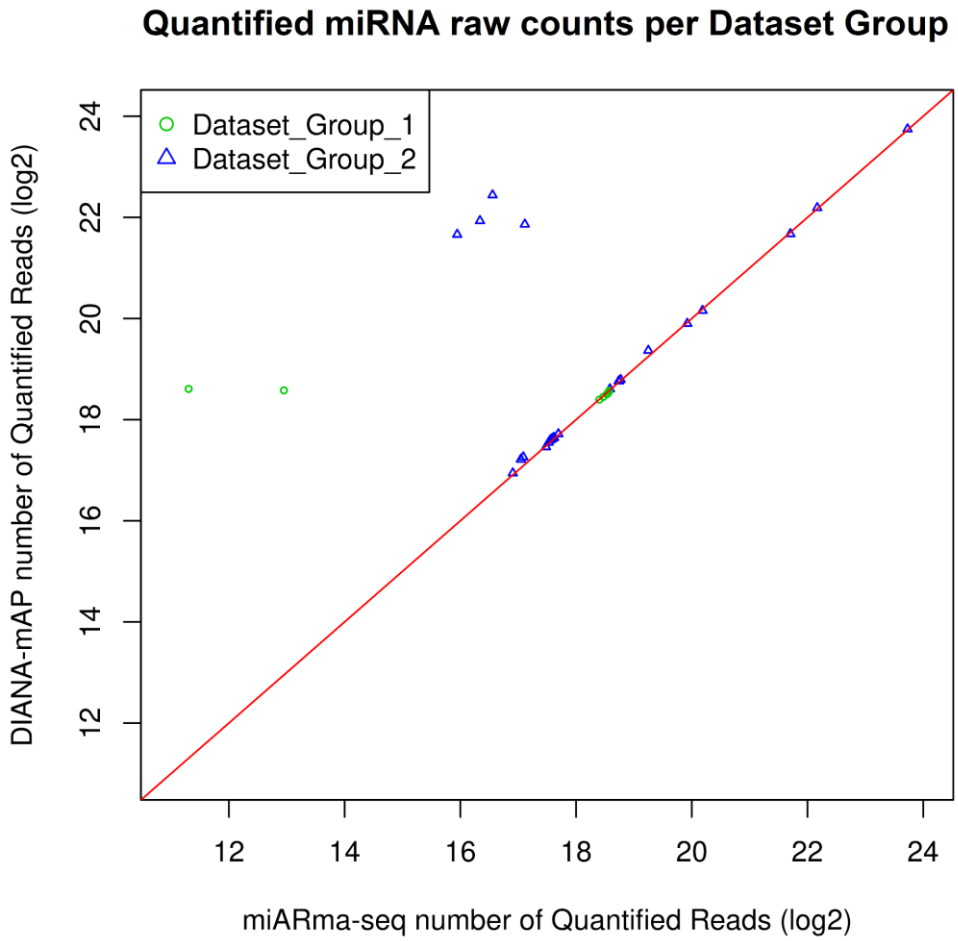

Figure 10. *Scatter plot of the quantified miRNA raw counts (Log_2-transformed) produced by DIANA-mAP and miARma-Seq tools by analysing both dataset groups* [63]

Furthermore, we conducted an evaluation of the reliability of our tool's quantification results by utilising an artificial sRNA-Seq dataset that we created for comparison with the Manatee quantification algorithm's published study [75]. This dataset was generated by employing

Monte Carlo random sampling methods based on real datasets. It consists of 778,072 simulated reads, excluding adapters, representing various known small RNA species such as miRNA (39.7%), rRNA (29.3%), tRNA (12.6%), and snoRNA (10.1%). A small fraction of reads also contains random modifications or mismatches. We analysed this sample using the specified parameters for both DIANA-mAP [63] and miARma-Seq [48]. Despite the presence of a large number of non-miRNA-related simulated reads and read replicates, DIANA-mAP successfully detected the absence of adapter sequences using default settings. The quantification results revealed that DIANA-mAP quantified 38.7% out of the 39.7% miRNAs included in the artificial dataset, compared to 33.5% for miARma-Seq (Figure 11).

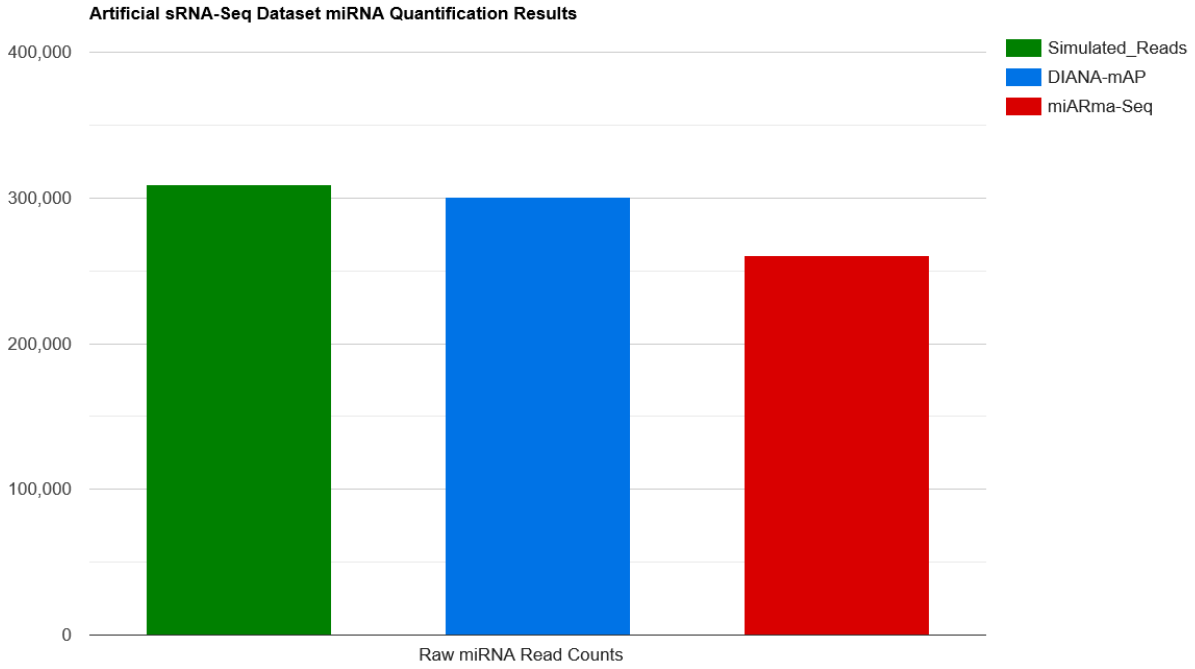

Figure 11. *Bar plot of the raw quantified miRNA results for DIANA-mAP and miARma-Seq on the artificial sRNA-Seq dataset* [63]

Furthermore, a correlation study of the raw miRNA expression values produced by both tools demonstrated a higher Pearson correlation between our tool's results and the simulated read counts, in comparison to miARma-Seq (Table 2).

	Pearson Correlation Coefficient	**Pearson p-value**
DIANA-mAP vs. Simulated Reads	0.9602398	$<2.2 \times 10^{-16}$

miARma-Seq vs. Simulated Reads	0.9376623	$<2.2 \times 10^{-16}$
DIANA-mAP vs. miARma-Seq	0.9231243	$<2.2 \times 10^{-16}$

Table 2. *Correlation study of raw miRNA expression results for DIANA-mAP and miARma-Seq on an artificial sRNA-Seq dataset* [63]

2.1.2. DIANA-RSeq

Overview

DIANA-RSeq is a fully automated computational RNA-Seq analysis pipeline that allows for RNA-Seq sample data quantification through a modular, scalable and easy to use multi-option tool. RNA-Seq analysis has been tackled for many years with multiple tools and software available for more or less specific analyses use cases. DIANA-RSeq pipeline aims to streamline and automate the RNA analysis process from raw RNA-Seq data to quantified expression results. The tool is composed of multiple standalone modules that comprise the data acquisition, pre-processing, alignment, quantification and quality control steps of the analysis. The alignment and the quantification modules provide multiple options on the software utilised (Figure 12), allowing for flexibility and choice in the balance between quality and required computational time.

Among the options, the analysis using the Salmon [76] software is suggested for well annotated organisms as a computationally cheap and reliable approach, while the use of STAR [77] aligner coupled with RSEM [78] is suggested for a more thorough but also slower analysis approach opted by major organisations such as ENCODE [69].

Snakemake has been utilised for DIANA-RSeq, allowing for scalable, standalone modules as well as eliminating the dependency installation through the use of Conda environments. Automated public dataset downloading is available from SRA [59]. DIANA-RSeq is freely available through Github. My contribution to the project was on the development and testing of the pipeline.

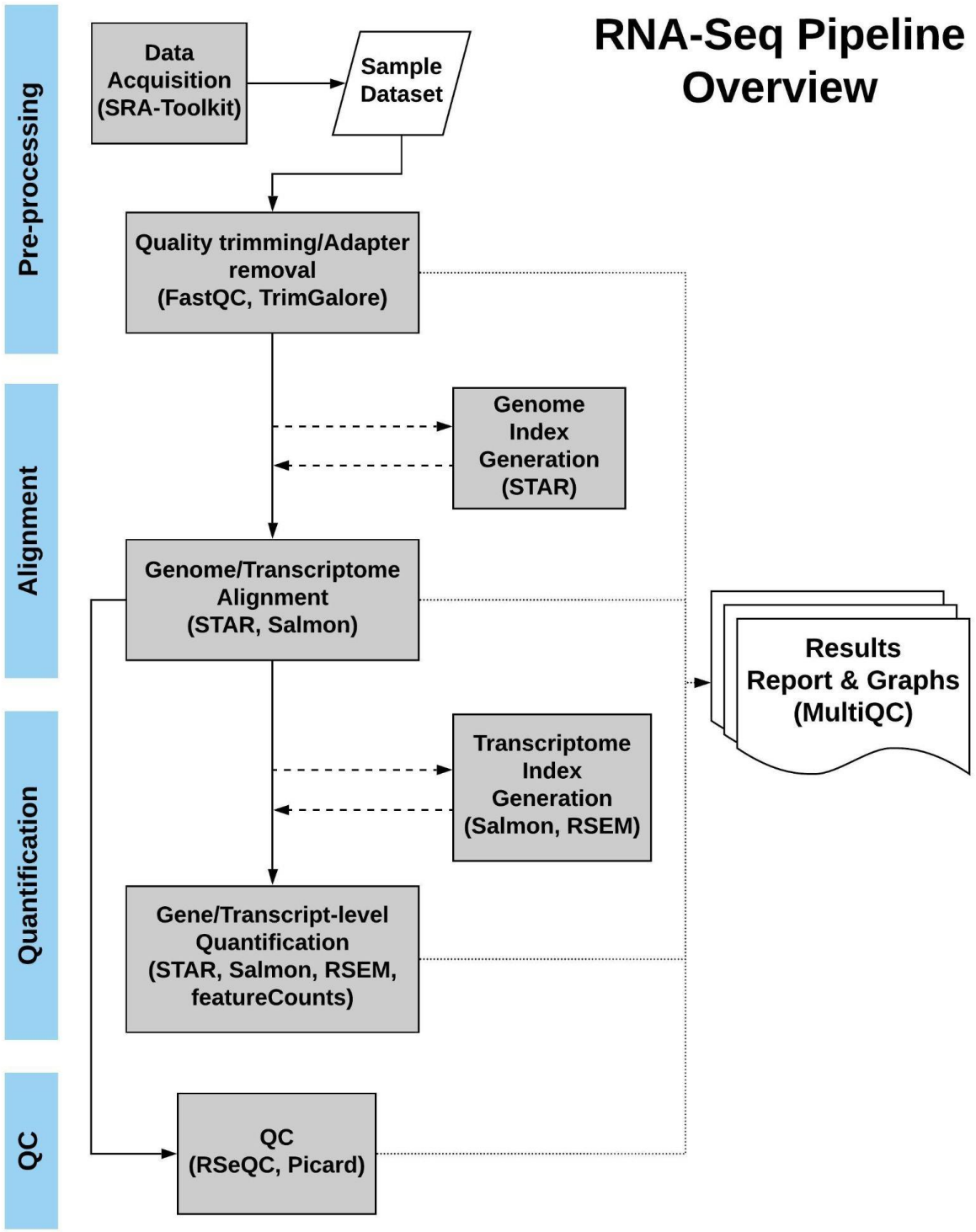

Figure 12. *The DIANA-RSeq analysis workflow and utilised tool options (figure created for the needs of the current thesis)*

Pre-processing Module

The Pre-processing module performs a quality check on the raw .fastq files and generates a summary report using MultiQC **[79]**. Following that, TrimGalore **[80]** is employed to trim a

provided adapter (or identify and trim one of the commonly used adapters) while also trim reads using a quality cutoff. Finally, a quality check is performed on the trimmed data and a MultiQC preprocessed summary report is generated.

Alignment Module

The Alignment module maps the input files to a provided genome reference file. It uses the STAR aligner software **[77]** and also generates a STAR index of the reference genome in case an index is not found in the appropriate directory provided in the configuration file. In order to improve the required computational time, the module utilises STAR's shared-genome-index option, loading the index on the computer's RAM only once thus avoiding index loading times for multiple alignments. Bam alignment files for both genome and transcriptome are generated and the genome aligned bam files are also sorted using the Samtools software **[81]**.

Quantification Module

The Quantification module infers the strandedness of each sample input and employs one of: Salmon **[76]**, RSEM **[78]**, featurecounts **[82]** and STAR **[77]** softwares to quantify the gene-level (and transcript-level for RSEM and Salmon) expression values of each sample. The quantifiers RSEM and Salmon also require a transcriptome index which will be automatically created, using the provided reference genome and annotation, if not found in the appropriate directory provided in the configuration file.

When using the Salmon option, if the Alignment module is included in the analysis, Salmon is using the alignment-based mode which quantifies the output .bam files from the Alignment module (in this mode a Salmon index is not required). If the Alignment module is not included, Salmon is used in mapping-mode which utilises Salmon's fast alignment followed by quantification (in this mode a Salmon transcriptome index is required and if not provided, a decoy-aware index is generated using the genome and the annotation files).

Quality Control Module

The Quality Control module provides a number of quality control information about each sample. Namely it uses the Picard **[83]** and RSeqQC **[84]** tools to provide alignment file (.bam) statistics, read duplication check and marking, read distribution, read inner distance, junction saturation and annotation.

Finally, apart from the utilised software results, MultiQC [79] is employed to create a summarising interactive report containing results and graphs from all the analysis steps, providing a thorough overview of the analysis with advanced filtering capabilities (Figure 13).

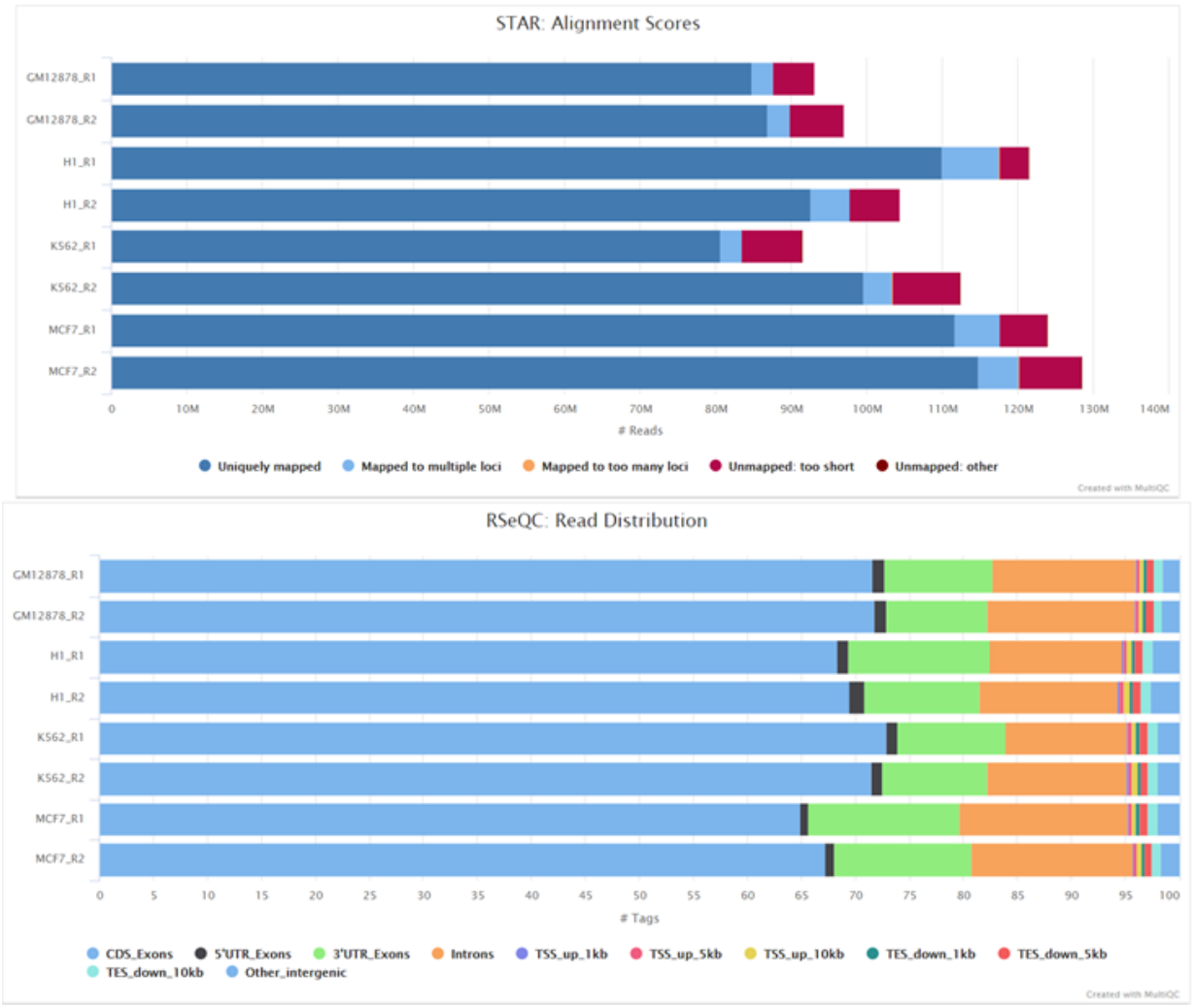

Figure 13. *DIANA-RSeq alignment and read distribution result graphs (figure created for the needs of the current thesis)*

2.2. Resources

Resources are available databases or web-servers that allow for their seamless use by users and provide information based on specific queries or input data. They are meant as an information provider and do not require the execution of any tool or algorithm by the user.

2.2.1. DIANA-miTED

Overview

The DIANA miRNA Tissue Expression Database (DIANA-miTED) [85] is the most comprehensive and systematic miRNA expression resource available to date, including expression values from over 15000 human small RNA-Seq samples from The Cancer Genome Atlas program (TCGA) [56] and the Sequence Read Archive (SRA) [59]. The usefulness of this expression atlas was enhanced by ensuring high-quality metadata. To achieve this, we carefully selected and organised information from SRA and TCGA sources. Our efforts resulted in a comprehensive and standardised collection that encompasses 199 tissues, 82 anatomical sublocations, 267 cell lines, and 261 diseases. Through miTED, users can easily access visualisations of expression and sample distributions based on their specific queries. Additionally, miTED provides study-wide diagrams and graphs that facilitate efficient exploration of the content. The tool also includes links to cutting-edge miRNA functional resources, making it an excellent starting point for expression retrieval, exploration, comparison, and downstream analysis. DIANA-miTED is freely accessible at http://www.microrna.gr/mited. My contribution to this project was in the development of the database along with the back- and frontend as well as the uniform bioinformatic analysis of the data included.

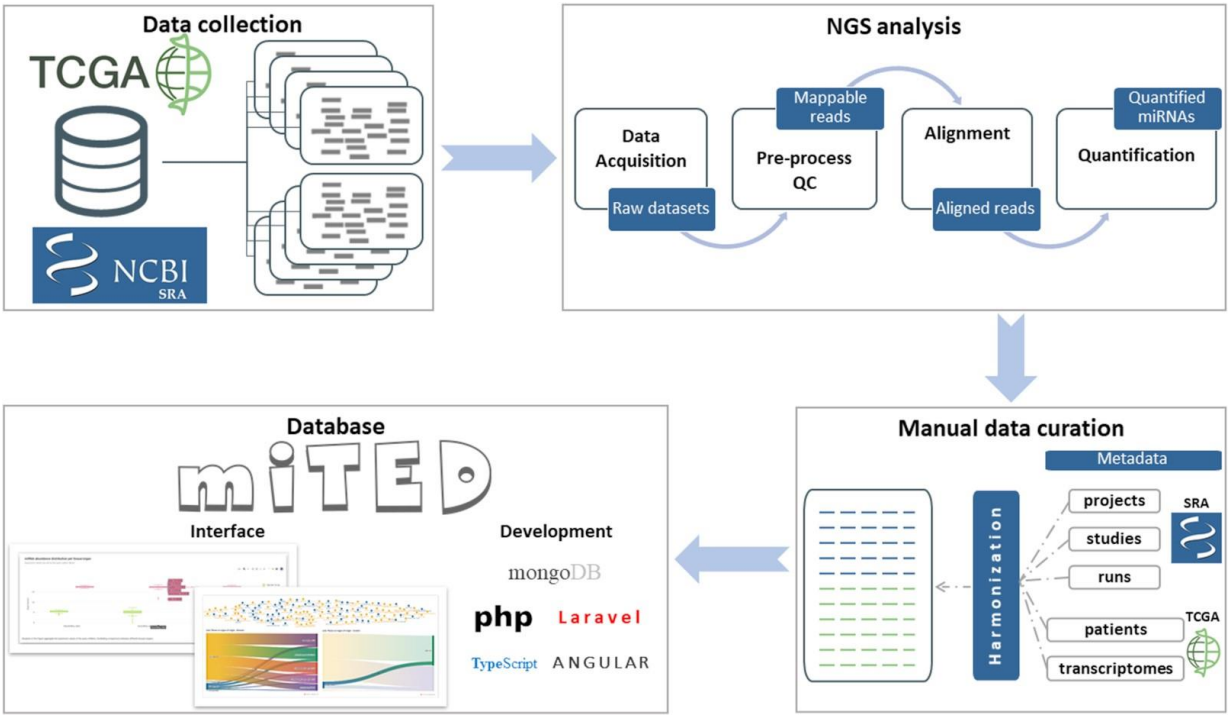

Figure 14. *DIANA-miTED development workflow* [85]

Data Collection and Analysis

Raw sRNA-Seq datasets originated from two resources, namely NCBI-SRA [59] and the TCGA project [56]. NCBI-SRA data were identified by selecting all entries with Library Strategy and organism fields as 'miRNA-Seq' and '*Homo sapiens*', respectively, and retrieved locally. Only datasets comprising tissue or cell line information metadata were included in the analysis. TCGA sRNA-Seq datasets were retrieved using the GDC Data Transfer Tool, along with the respective transcriptomic metadata from the GDC data portal, while all available sample and patient metadata were retrieved. Only sRNA-seq TCGA datasets annotated with both patient and transcriptomic metadata were included in the study.

Both NCBI-SRA and TCGA-derived metadata information were manually curated in order to deliver to users a coherent and standardised set. The resulting collection offers the following metadata information. 'Sample_ID' contains the sample's identification code in SRA (e.g. SRR1774098) or TCGA (e.g. TCGA-P8-A5KD-11A-11R-A35M-13). 'Collection' holds the origin collection of the sample (SRA or TCGA). 'Project_ID' contains the sample's project identification code in NCBI-SRA (e.g. PRJDB2675) or TCGA (e.g. TCGA-PCPG). All entries in miTED mandatorily contain the aforementioned information. Moreover, each entry must either feature 'Tissue or organ of origin' or 'Cell line' information. Depending on the initial metadata, 'Tissue subregion' and 'Disease' information is provided.

All datasets were analysed with the well-defined sRNA-Seq analysis pipeline DIANA-mAP [63], following strict quality controls, resulting in a collection of 4142 NCBI-SRA and 11 041 TCGA analysed datasets. Raw datasets were quality checked allowing a minimum Phred score of 10, adapters were inferred when not provided and trimmed using an 18 bp minimum allowed read length after trimming. Reads were subsequently aligned to the GRCh38 human genome assembly, allowing up to five multi-maps per read. Finally, the aligned reads were quantified towards the miRBase [67] v22 hairpin and mature human miRNA forms, allowing one mismatch on the precursor during alignment. The remaining DIANA-mAP and external tools' parameters were set to their default values.

When analysing the NCBI-SRA datasets, adapter information from metadata was used when available, although it was scarce and therefore adapter inference was performed whenever required. sRNA-Seq datasets often have low mapping rates, with less than half of the reads assigned to miRNAs due to factors like sample contamination. Despite this, the quantification results can still be considered reliable [86]. In our study, a large number of samples were analysed, but adapter information was missing in many cases. To ensure accuracy and avoid erroneous results due to possibly incorrect adapter inference, we only included samples where at least 50% of the pre-processed reads were assigned to known miRNAs. The same criteria were applied to the analysis of TCGA datasets, resulting in an average rate of 79.2% [76.6-81.9, 95% confidence interval] of pre-processed reads assigned to known miRNAs. Read counts and read per million (RPM) units were obtained directly from DIANA-mAP results, while log2(RPM + 1) values were calculated specifically for visualisation purposes.

Database Implementation

DIANA-miTED is an NoSQL database that follows the MVC architecture and is hosted on Apache HTTP Server 2.4. The back-end, responsible for data access, utilises MongoDB (https://www.mongodb.com/) and the PHP framework Laravel 8 (https://laravel.com/) (PHP 7.2). For the front-end, the presentation layer is created using Angular 9.1 (https://angular.io/) and incorporates the Angular Material UI library (https://material.angular.io/). The data of miTED is stored in collections within the NoSQL database, and Laravel facilitates the connection to these collections for storing and retrieving data. Regarding the presentation layer, database statistics are displayed using Chart JS (https://www.chartjs.org/) and the Plotly JavaScript Open Source Graphing Library (https://plotly.com/javascript/), while Flourish (https://flourish.studio/) is employed for more intricate visualisations.

By analysing over 120 billion reads obtained from 15,183 sRNA-seq datasets, we extracted a total of more than 120 million expression values. These expression values were calculated as read counts, reads per million (RPM), and log2(RPM). Out of the datasets analysed, approximately 27.3% (4,142 datasets) were obtained from NCBI-SRA [59], while the remaining 72.7% (11,041 datasets) were sourced from TCGA [56].

The database contains a wide range of datasets, including 12,400 datasets derived from disease samples representing 261 different human pathologies. Additionally, there are 1,386 datasets available from healthy samples. Within miTED, there are 14,146 samples originating from 199 distinct tissues/organs, and 1,037 samples from 267 different cell lines.

In terms of gender distribution, miTED maintains a balanced representation, with 6,751 samples derived from females, 6,123 samples from males, and 2,309 entries that are unannotated or do not specify the gender.

Interface and Functionality

To facilitate easy access and analysis of this extensive collection, an intuitive online graphical user interface was developed. This interface allows users to search, browse, and perform meta-analysis without requiring specialised bioinformatics support or expertise. DIANA-miTED offers three primary query pages accessible from the Querying DB top menu. The Multi-query page enables users to explore and compare the expression of specific miRNAs in different tissues or cell lines. On the Top-miRNAs page, users can obtain a list of the most highly expressed miRNAs in a particular tissue or cell line. Lastly, the Top-sites page provides information on the tissues or cell lines where a specific query miRNA exhibits the highest expression levels. All the generated results and plots can be conveniently downloaded without the need for login, application, or verification procedures. Dedicated download buttons are provided to ensure easy access to the desired data and visualisations.

On the Multi-query page, users have the capability to perform queries, retrieve, and compare the expression of one or multiple miRNAs in tissues or cell lines. The page includes dedicated search boxes that allow users to conduct free-text searches and select specific tissues, cell lines, and miRNAs (Figure 15A). The Multi-query form offers several options to refine the search. Users can specify particular diseases, choose to include results from either the SRA or TCGA

data collections, filter data based on health status (such as 'Healthy' or 'Disease'), and select the desired expression unit, which can be read counts, RPM, or log2(RPM).

The results from the query are presented in three distinct sections. The first section focuses on visualising the retrieved results. Grouped boxplots enable users to compare miRNA abundance in specific tissues or diseases (Figure 15C). Additionally, sample distributions are depicted using a Sankey diagram, illustrating relationships between tissues and diseases, and pie charts that provide information on gender, data collection, and health status (Figure 15D).

Figure 15. *DIANA-miTED multi-query page interface* [85]

The second section facilitates the integration of miTED results with related DIANA resources, including various tools and databases, for each miRNA. Hyperlinks are provided for each input

miRNA, connecting users to DIANA-microT-CDS [87], a web server for predicted miRNA targets, DIANA-Tarbase v.8 [88], a reference database of experimentally supported miRNA targets, DIANA-LncBase v.3 [89], a repository of experimentally supported miRNA targets on long non-coding RNAs, and DIANA-miRPath v.3 [90], a dedicated web server for assessing miRNA regulatory roles and identifying controlled pathways. Finally, the third section presents a data table that includes sample metadata and the expression levels of the miRNAs requested by the user, as described in the Materials and Methods section (Figure 15B).

The Top-miRNAs page serves as the second query page within the miTED resource. Users can utilise this page to search for the most highly expressed miRNAs in a specific tissue or cell line. The results displayed include a data table that presents the expression levels of all miRNAs in descending order, as well as a bar chart visualising the top expressed miRNAs in the selected tissue or cell line.

The Top-sites page is specifically designed for retrieving information about the tissues or cell lines where a particular miRNA exhibits the highest abundance. Similar to the Top-miRNAs page, the results on this page consist of a table that lists the expression values of tissues or cell lines in descending order, along with a bar chart illustrating the top tissues or cell lines where the queried miRNA is expressed the most.

Within DIANA-miTED, users can access three visualisation pages through the Visualisations menu. The first page, titled 'Tissue - Subregions | Graph Network,' presents an interactive graph network that showcases the relationships between the tissue or organ of origin and its corresponding subregions (Figure 16A). Users have the ability to highlight and manipulate nodes, allowing them to explore the interconnectedness between different elements within the graph.

The 'TCGA Projects Exploration' page offers Sankey diagrams specifically designed for investigating the relationships between tissues and diseases, as well as tissues and gender within the TCGA datasets (Figure 16B). These Sankey diagrams are interactive, enabling users to explore the distribution of samples within each category.

Lastly, the 'DB statistics' page provides additional graphs that offer an overview of the overall content within the database, serving as supplementary information for users to gain insights into the database's composition and characteristics.

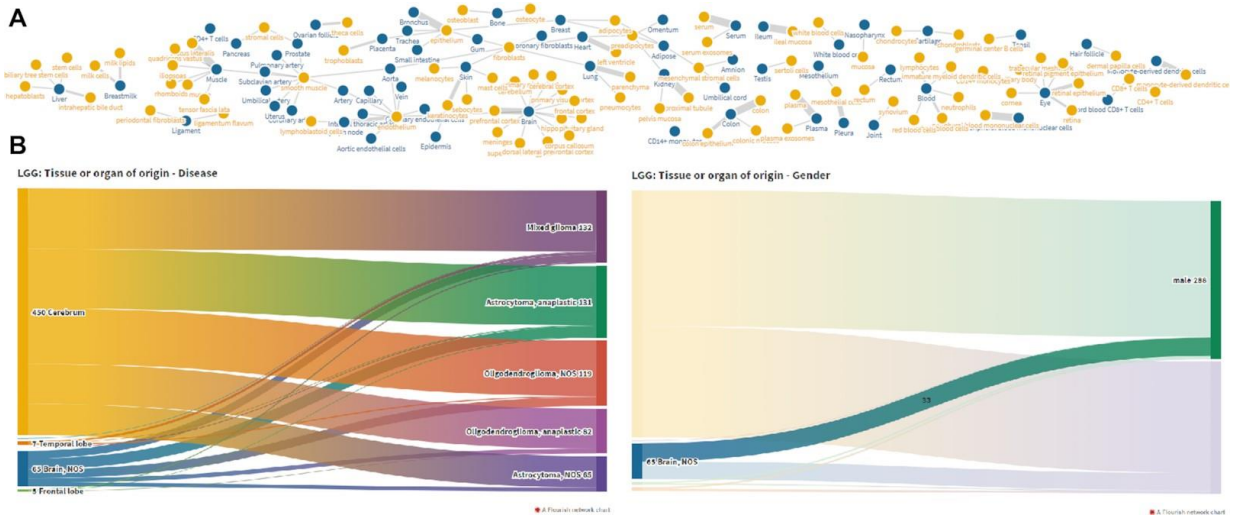

Figure 16. *DIANA-miTED visualisations* [85]

2.2.2. PlasmiR

Overview

Recently, the discovery of microRNAs (miRNAs) in detectable and distinct amounts in the human circulatory system has opened up the intriguing possibility of utilising them as minimally invasive biomarkers. Extensive research in various disease contexts, including cancer, aims to investigate the protective or pathogenic functions of dysregulated miRNAs and explore their potential as biomarkers. However, there is currently no specialised resource that compiles experimentally validated circulating miRNA biomarkers.

To address this gap, we conducted a thorough search of the existing literature to identify articles that assessed the diagnostic or prognostic roles of miRNAs in blood, serum, or plasma samples. We carefully scrutinised these articles, excluding those that lacked sufficient experimental documentation or did not employ biomarker assessment methods. From over 200 biomedical articles, we gathered crucial meta-information including cohort sizes, inclusion-exclusion criteria, disease/healthy confirmation methods, and quantification details. Additionally, we systematically characterised the miRNAs and diseases using reference resources. The culmination of our efforts is an online database called plasmiR [91], which presents a collection of circulating miRNA biomarkers. This database encompasses 1,021 entries, involving 251 miRNAs and 112 diseases. Notably, more than half of the entries in plasmiR pertain to cancerous and neoplastic conditions, with 183 of them (32%) describing

prognostic associations. plasmiR offers user-friendly features such as smart queries, emphasising visualisation and exploratory modes to aid researchers in their investigations.

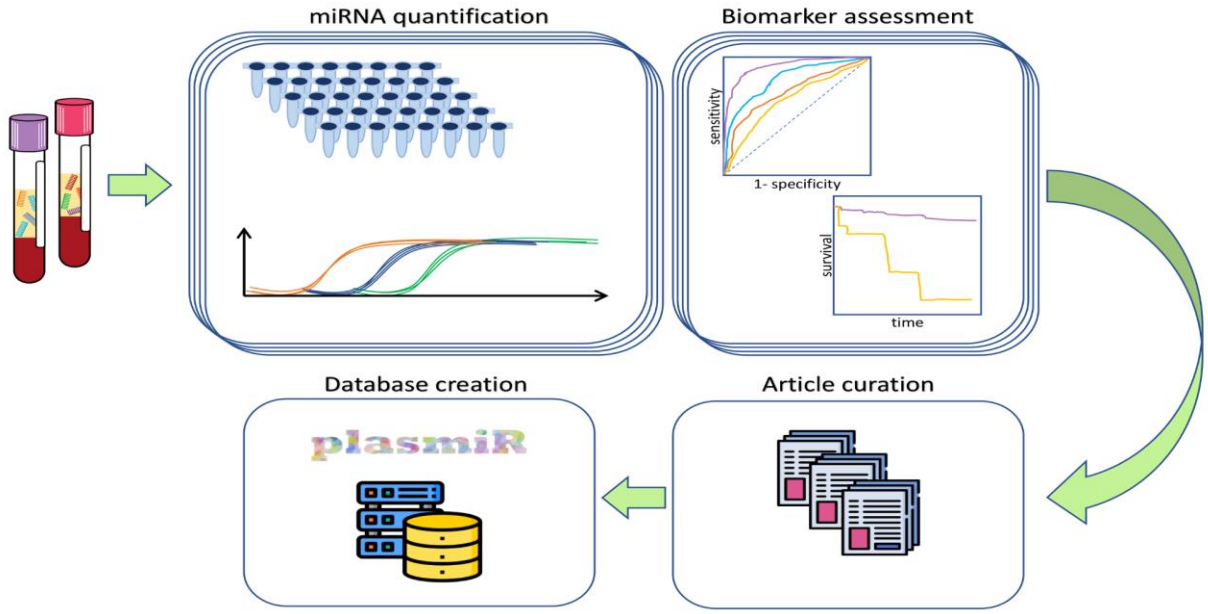

Figure 17. *The plasmiR creation workflow* [91]

In summary, plasmiR fills a crucial void by providing a comprehensive and curated collection of circulating miRNA biomarkers, derived from a rigorous analysis of relevant scientific literature. It serves as a valuable resource for researchers seeking to explore the potential of miRNAs as diagnostic or prognostic indicators. The resource is freely available at http://www.microrna.gr/plasmir. My contribution to this project was the development of the database along with the back- and frontend.

Data Collection and Curation

A systematic approach was taken to compile a database of miRNA-related publications. PubMed and PubMed Central were initially searched using relevant keywords, followed by a filtering process to select appropriate articles based on titles and abstracts. Additional articles referenced within review papers were also considered. This process resulted in a set of 411 candidate publications spanning from 2003 to 2021. A template sheet was created to curate primary data and metadata, including fields for disease, miRNA, biomarker, and study information.

The purpose of each entry in the database, whether diagnostic or prognostic, was clearly distinguished. Separate entries were created for studies with repeated assessments or those

exploring different aspects of miRNA roles. Free-text fields were used for disease and prognostic outcome, while miRNA names were validated using the miRBase reference database [67]. Discrepancies were documented, and quality control checks were performed by independent curators.

Integration with relevant resources such as RNAcentral [92], MirGeneDB 2.0 [93] and disease databases like the Comparative Toxicogenomics Database (CTD) [94] ensured accurate annotations. Hyperlinks were created to establish active connections with various resources, including miRBase [67], RNAcentral, CTD, Medical Subject Headings (MeSH) Disease [95], Disease Ontology [96], Online Mendelian Inheritance in Man (OMIM) [97], PubMed, DIANA-TarBase v8.0 [88], and DIANA-miRPath v3.0 [90]. This facilitated easy access to additional information and related resources for each entry in the database.

Data Statistics

The curation process resulted in 1021 entries from 204 research articles. The plasmiR resource [91] provided comprehensive information on 251 circulating miRNAs and 112 systemic disease names. A total of 594 unique miRNA-disease pairs were established, representing distinct combinations of annotated miRNAs and diseases. Around 30% of the miRNAs (80 out of 251) had both diagnostic and prognostic capabilities. The diseases covered a range of categories, including cancer, cardiovascular conditions, neurological disorders, metabolic diseases, and various other pathological conditions. The largest number of entries (565) and miRNAs (149) were associated with cancers and neoplasms. The "Other" category included diverse conditions such as infection-related diseases, pregnancy complications, liver diseases, hormone deficiencies, and autoimmune diseases.

The majority of potential biomarker miRNAs, except for neurological disorders, are up-regulated in the studied disease states compared to healthy controls. Most miRNAs demonstrate a limited range of biomarker potential within specific disease categories, associated with one or a few diseases. For diagnostic entries, the majority of miRNAs in each disease category have up to three biomarker entries per miRNA, ranging from 77% in cancer to 96% in the "Other" disease category. A similar pattern applies to miRNAs with prognostic capacity in cancer. Some miRNAs associated with cancer have both diagnostic and prognostic roles, accounting for a significant portion of diagnostic cancer miRNAs. There is also evidence of miRNA overlap across different disease categories, with certain miRNAs shared between cancers and cardiovascular diseases.

In terms of evaluating the biomarker value, the majority of diagnostic entries have been

validated using ROC analysis, while various techniques such as the log-rank test, Kaplan-Meier analysis, odds ratios, and Cox Regression analysis were employed for prognostic entries. A significant number of entries assessing prognostic value specifically focus on patient survival, including various survival metrics.

Database Implementation

The MVC architecture was employed to build an SQL database, which was hosted on Apache HTTP server 2.4. The back-end components consisted of PostgreSQL v12.6 (https://www.postgresql.org/) and the PHP framework Laravel 8 (https://laravel.com/) using PHP 7.2. For the front-end development, Angular 9.1 (https://angular.io/) and the Angular Material UI library (https://material.angular.io/) were used. The plasmiR data is stored in SQL tables, and the retrieval of data from these tables is managed by Laravel. The database design was focused on reducing the redundancy of the data in the tables while providing flexibility in querying and data presentation based on results (Figure 18).

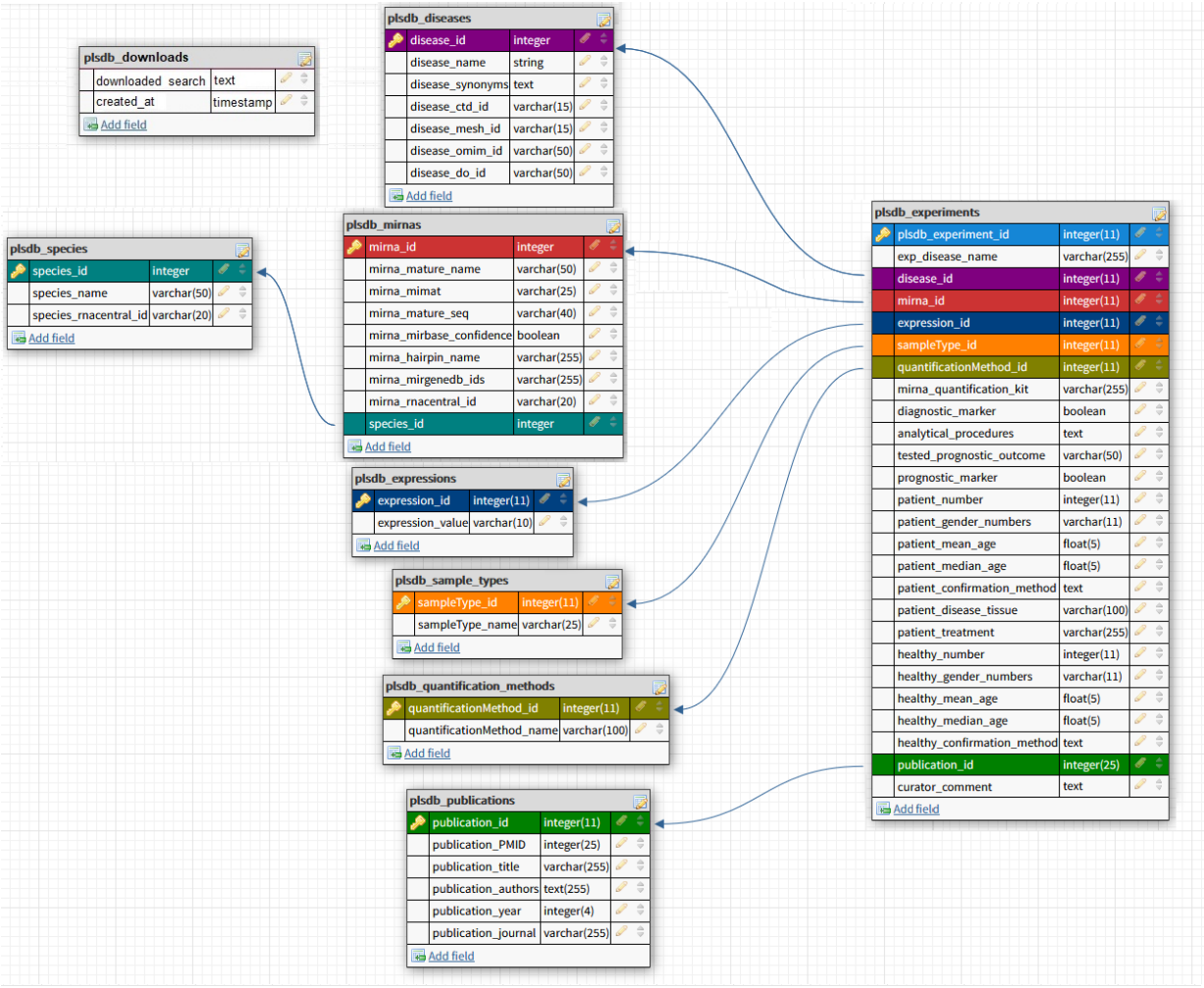

Figure 18. *The plasmiR database schema (figure created for the needs of the current thesis)*

To present database statistics and result-specific visualisations, the presentation layer incorporates Chart JS (https://www.chartjs.org/) and the Plotly JavaScript Open Source Graphing Library (https://plotly.com/javascript/). Finally, the Flourish online application (https://flourish.studio/) is employed to generate the interactive miRNA-disease network graphs, enabling users to explore the connections between miRNAs and diseases in an intuitive manner.

Interface and Functionality

plasmiR allows users to perform queries by providing mature miRNA names, systemic disease names, or a combination of both. Various filtering options are available to refine the search, including miRNA expression direction, sample/biomarker type, cohort age range, and minimum cohort size. Drop-down menus have been implemented to prevent typographical errors in the queries.

The primary information provided in plasmiR entries includes the miRNA-disease pair, the sample type (plasma, serum, or blood), the biomarker type, and, in prognostic entries, the assessed outcome (Figure 19a). Users can expand the view of entries of interest to access comprehensive metadata, including experimental and statistical details, relevant publications, curator comments, and links to external resources. By clicking on links to DIANA-TarBase and DIANA-miRPath, users can explore experimentally supported miRNA targets and perform in silico functional analysis of specific miRNAs. This functionality is particularly valuable for investigating the functional implications of changes in circulating miRNA abundance associated with disease-related events, such as tissue injury or tumor metastasis. To facilitate downstream analyses, plasmiR provides the option to download query results locally in a tab-delimited format, in addition to offering direct links to other online resources.

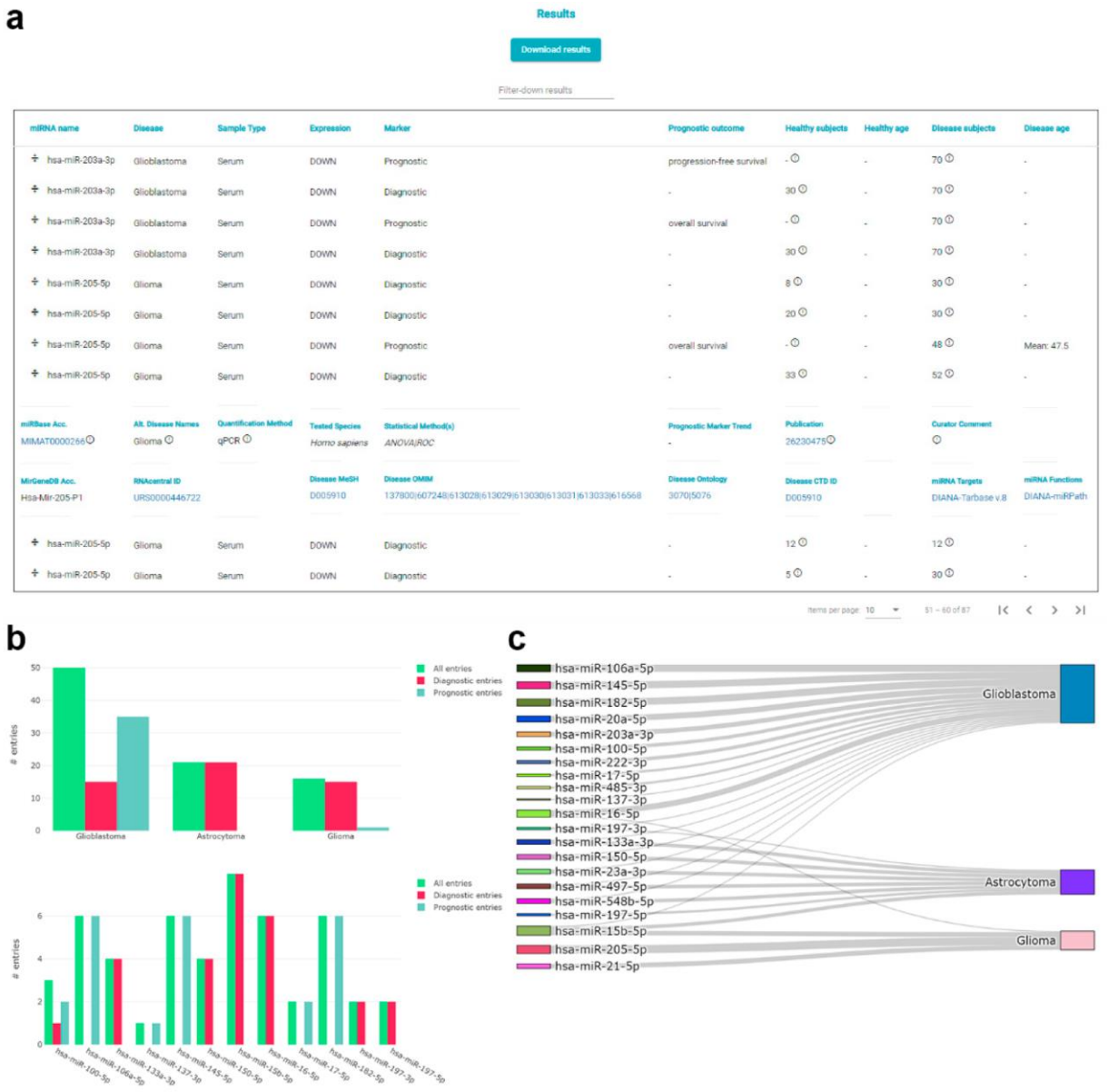

Figure 19. *The plasmiR results interface* [91]

Result-specific plots are also generated for each query, which can be saved locally. These plots include two separate bar plots, one displaying results per disease and the other per miRNA (Figure 19b). Additionally, a Sankey plot (Figure 18c) is utilised to visualise the relationships between biomarkers and diseases, such as miRNAs that serve as biomarkers in multiple diseases versus unique miRNA-disease pairs.

Additionally, Statistics and Visualisations pages are provided that give database-wide insights and assist in guiding specific queries. In the Statistics page, horizontal bar plots present the top miRNAs and diseases in terms of the absolute number of diagnostic and prognostic entries. On the Visualisations page, four interactive network graphs are available for the main disease

categories covered in plasmiR, including cancers, neurological conditions, cardiovascular disorders, and metabolic diseases. Nodes represent miRNAs and diseases, with color-coded miRNAs indicating their diagnostic and/or prognostic potential. A help section is also included to provide a comprehensive description of each component within the plasmiR resource, facilitating easy navigation through its content.

2.2.3. DIANA-microT 2023 web-server

Overview

DIANA-microT-CDS, an advanced miRNA target prediction algorithm, has been serving the scientific community since 2012 [98]. It was among the pioneering algorithms that accurately predicted miRNA binding sites in both the 3′ Untranslated Region (3′-UTR) and the coding sequence (CDS) of transcripts, achieving improved performance. The latest version, DIANA-microT 2023 (available at www.microrna.gr/microt_webserver/), offers a significantly updated collection of interactions [99].

In the development of this last web server version, the DIANA-microT-CDS algorithm was executed using annotation data from Ensembl v102 [100], miRBase 22.1 [67], and MirGeneDB 2.1 [93] were utilised for the first time. This integration resulted in over 83 million interactions for various species, including human, mouse, rat, chicken, fly, and worm. Furthermore, this version provides predictions of interactions between miRNAs encoded by 20 viruses and host transcripts from human, mouse, and chicken species.

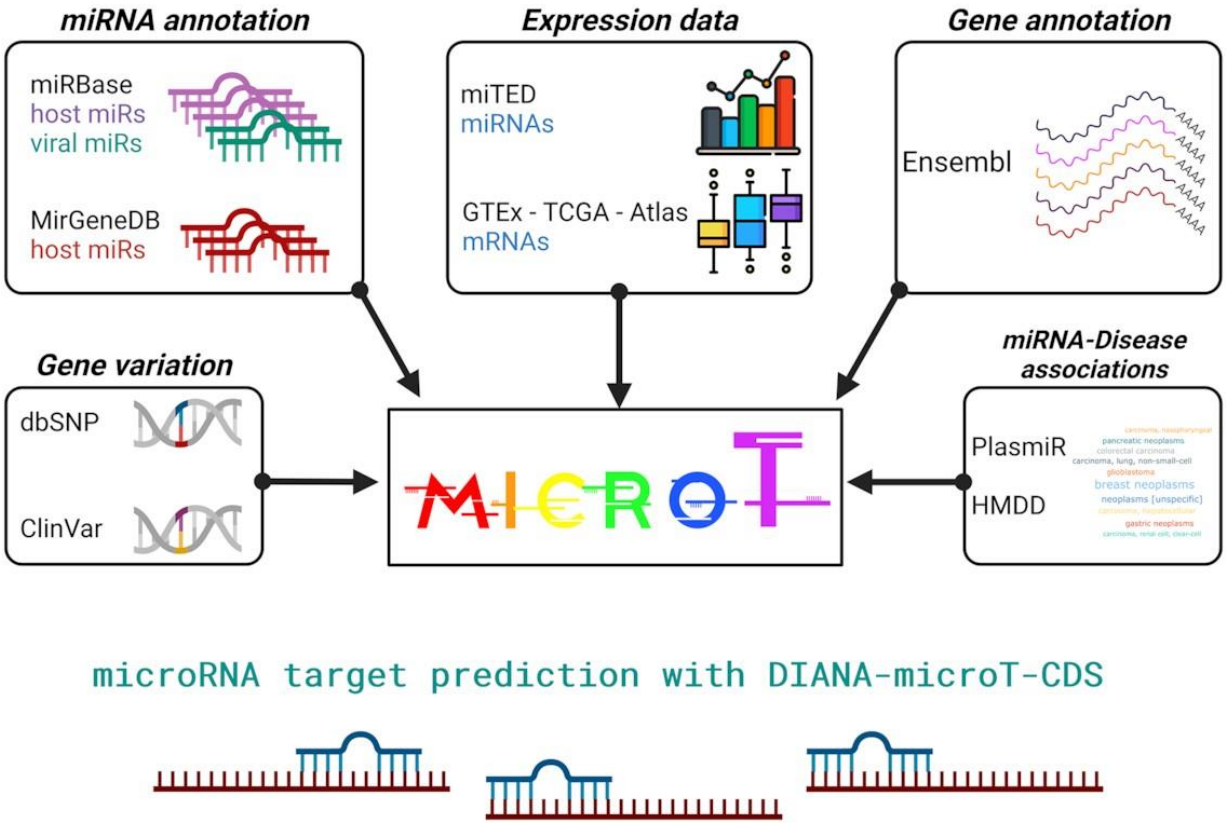

Figure 20. *The DIANA-microT 2023 web server overview* [99]

The DIANA-microT 2023 web server incorporates numerous resources such as DIANA-TarBase [88], plasmiR [91], HMDD [61], UCSC [101], dbSNP [102], ClinVar [103], and miRNA/gene abundance values for 369 distinct cell lines and tissues. The server interface has been redesigned, offering users smart filtering options, the ability to identify abundance patterns of interest, pinpoint known SNPs located on binding sites, and access miRNA-disease information. Finally, the contents of the DIANA-microT 2023 webserver are freely accessible, and users can also download the data locally without the need for any login credentials. My main contribution to this project was the full stack development of the database.

Data Generation and Collection

For transcript annotations, Ensembl v102 [100] served as the reference, specifically selecting protein-coding genes located on the standard chromosomes. To obtain miRNA sequences and annotations, miRBase v22.1 [67] and MirGeneDB 2.1 [93] resources were utilised. The corresponding genome assemblies (e.g., hg38, 30-way for H. sapiens; mm10, 60-way for M. musculus; rn6, 30-way for R. norvegicus; galGal6, 77-way for G. gallus; dm6, 27-way for D. melanogaster; and WBcel235/ce11, 26-way for C. elegans) provided multiple sequence

alignments, acquired through the UCSC genome browser [101].

For the CDS and UTR regions, secondary structure predictions were performed using Sfold [104]. microT-CDS was executed separately for each species, incorporating the miRBase and MirGeneDB annotations. This process resulted in a comprehensive collection of approximately 83.9 million interactions (with 5.9 million interactions meeting the default microT score threshold of 0.7). Additionally, predictions of v-miRNA sequences were performed for human (viruses:EBV, KSHV, HCMV, HSV1, HSV2, HIV1, HHV6B, MCPV, JCV, BKV, TTV, SFV, HBV and SV40), mouse (viruses: MGHV and MCMV), and chicken (viruses: MDV2, MDV1, HVT and ILTV). The v-miRNA entries provided encompass approximately 2.5 million interactions, with 185,460 interactions having a microT score of at least 0.7.

Conservation information was obtained from UCSC [101] while experimentally verified interactions from DIANA-TarBase v8.0 [88] and TargetScan [105] predictions were obtained and matched to DIANA-microT-CDS predictions. Gene expression data for human and mouse were obtained from GTEx [106], TCGA [56] and the Sollner *et al.* mouse expression atlas [107], resulting in 99 distinct tissue states. Similarly, summarised miRNA abundance estimates were obtained through DIANA-miTED [85] as Reads-Per-Million (RPM) values. They correspond to 270 cell-line/tissue states in humans (miRBase [67] annotation only; non-viral miRNAs). Variant annotation SNP VCF files were retrieved from dbSNP v151 [102] and ClinVar [103]. Finally, causal associations of miRNAs against diseases were retrieved from HMDD 3.2 [61] while circulating miRNA biomarker information was derived from *plasmiR* [91].

Database Implementation

The microT 2023 web server is built on the Model-View-Controller (MVC) software architecture and utilises the Laravel 8 PHP framework as its foundation. It communicates with the Angular-based frontend through a RESTful interface. The webserver is hosted on an Apache 2.4 HTTP server, and its data is stored in a relational database managed by a PostgreSQL 11.8 server.

The back-end logic, including the connection to the PostgreSQL server for data storage and retrieval, is handled by the PHP framework Laravel 8 (https://laravel.com/) running on PHP 7.2. The front-end of the web server is designed as a one-page website using Angular 14 (https://angular.io/). It incorporates the Angular Material UI library(https://material.angular.io/) and the ngx-bootstrap framework (https://valor-software.com/ngx-bootstrap) for its visual and functional components. Due to the vast number of interactions, a thorough database design

process took place in order to avoid duplications in database tables as well as to optimise the number of table joins required for the queries (Figure 21).

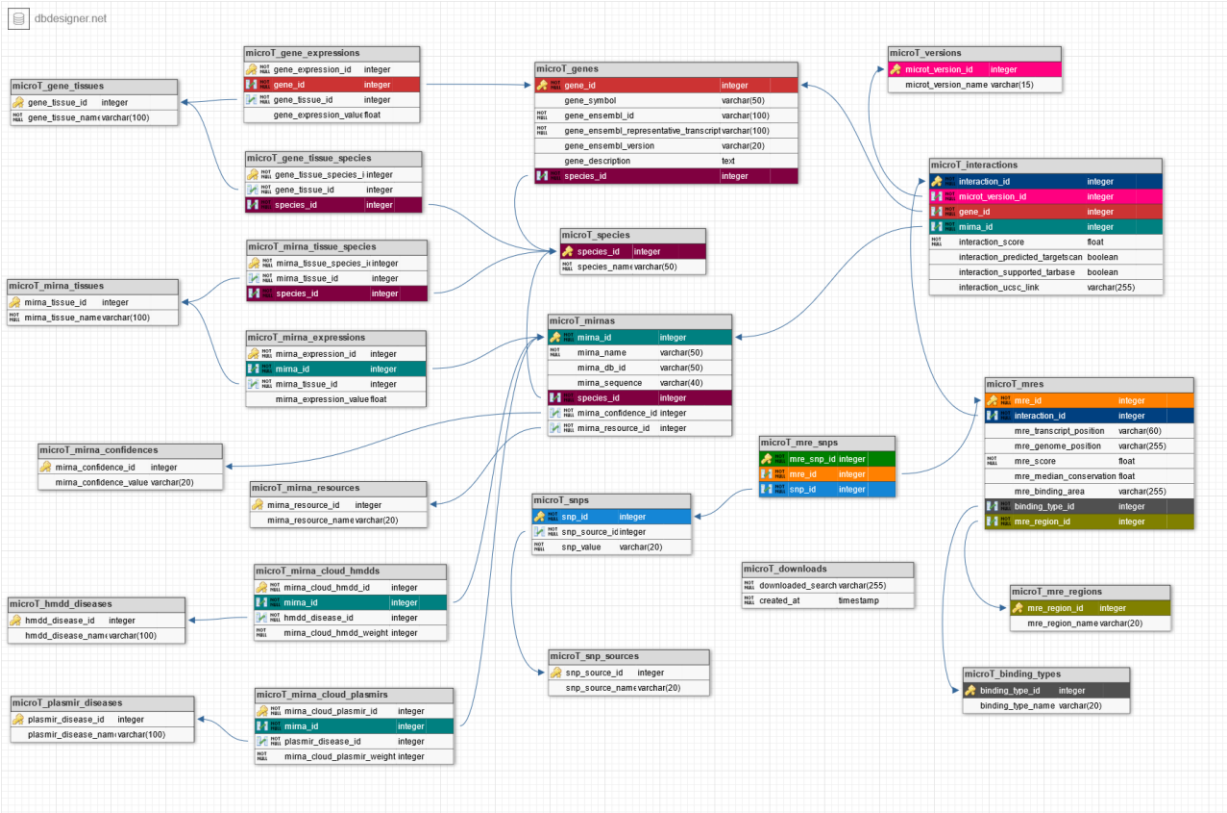

Figure 21. *The DIANA-microT 2023 web server database schema (figure created for the needs of the current thesis)*

To present database statistics, the web server utilises the Chart JS library (https://www.chartjs.org/). Additionally, AnyChart (https://www.anychart.com/) is employed to generate word-cloud visualisations that display the frequency of diseases associated with specific miRNAs. These visualisations include active hyperlinks that direct users to the HMDD [61] and plasmiR [91] servers.

Interface and Functionality

The previous version of DIANA-microT had a simple interface for miRNA target predictions. It consisted of a miRNA/gene input service and a manual filter to set the prediction score threshold. This user-friendly approach is carried forward in the interface of DIANA-microT 2023 web server, making it convenient to retrieve desired predictions quickly (Figure 22A). The basic menu requires selecting the target species, choosing a miRNA annotation source, and providing one or more miRNAs. However, for more advanced queries, users can expand a

supplementary menu providing multiple filter options (Figure 22B).

In addition to setting the minimum prediction score, the webserver offers options to generate interactions based on highly reliable miRNA annotations, outputting only results related to the coding sequence (CDS) or 3′-UTR, or including only interactions supported by TarBase and/or TargetScan. Notably, there are two dedicated controls that allow users to select specific tissues and cell lines, providing abundance information in the annotated results.

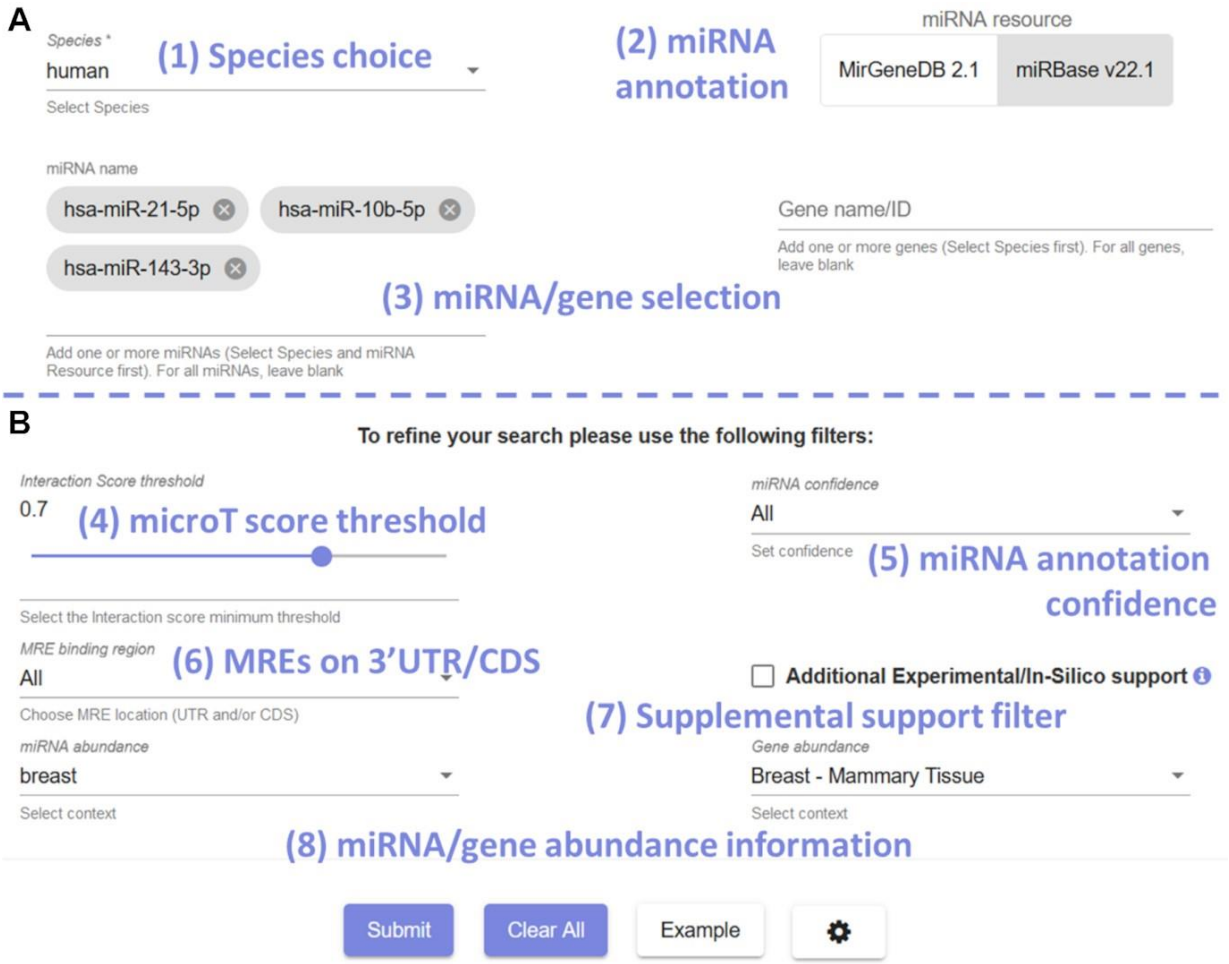

Figure 22. *The DIANA-microT 2023 query interface* [99]

The interactions table generated has been redesigned to present all relevant information in a user-friendly hierarchical structure (Figure 23). The primary layer of the table displays details at the interaction level. Each miRNA-gene pair is accompanied by an interaction score and notation indicating predicted or experimental support. Additionally, a link is provided to a dedicated UCSC track that showcases the positions of all interaction-specific MREs (microRNA response elements) and z-scores of abundance metrics if specific expression contexts have been selected for genes and/or miRNAs.

Abundance metrics offer context-specific experimental support. For example, in the given example, hsa-miR-143-3p shows high expression compared to other miRNAs (6.3 standard deviations higher than the miRNAs' mean), while PSG4 appears to be moderately repressed (1.25 standard deviations lower than the genes' mean). Clicking on gene identifiers and miRNA names reveals further details at the gene and miRNA levels, including causal associations and disease-clouds associated with each miRNA.

Expanding an entry of interest provides supplemental information at the MRE level. Basic MRE details include the region (3′-UTR or CDS), MRE coordinates on the respective transcript or genome, primary binding type, and MRE score. Additionally, the average phastCons conservation score of the predicted MRE is provided. Pop-up buttons are available to retrieve a schematic of the MRE binding area and obtain information about any SNPs that overlap with the MRE sequence.

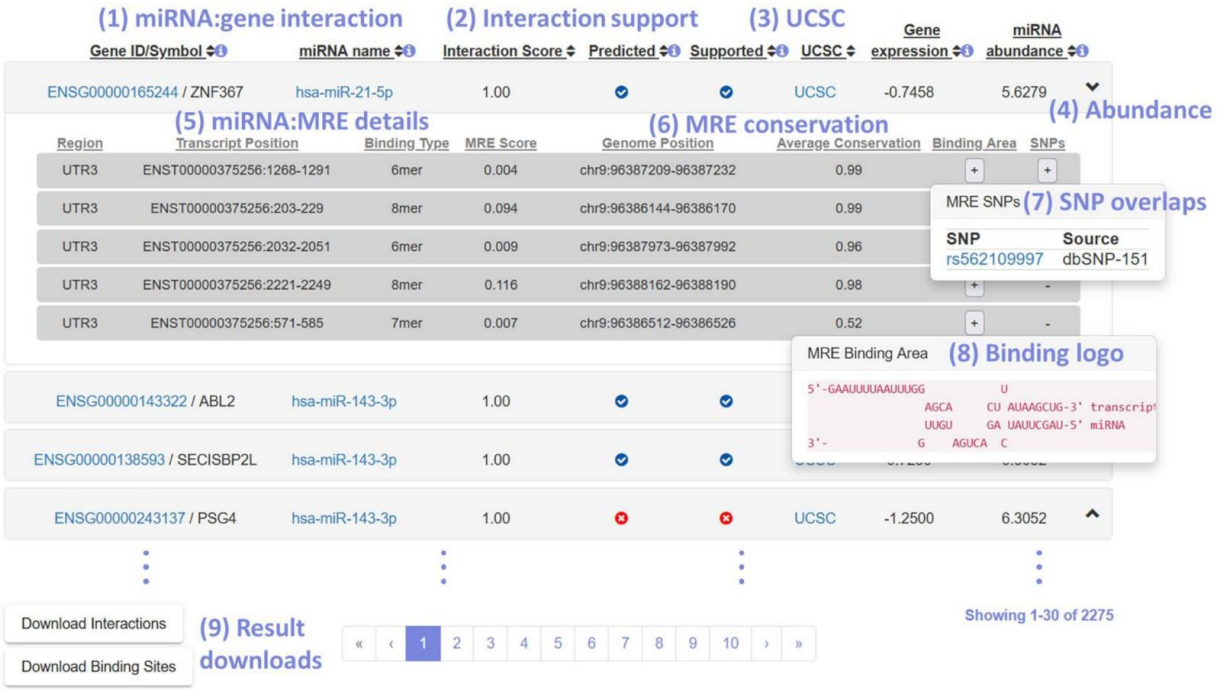

Figure 23. *The DIANA-microT 2023 output format* [99]

2.2.4. Peryton

Overview

Peryton [108] is a database that offers comprehensive support for microbe-disease associations based on experimental evidence. The initial version of Peryton presents a unique resource

featuring over 7,900 entries that link 43 diseases with 1,396 microorganisms. The content of Peryton is meticulously curated from biomedical articles, ensuring reliable and accurate information. Diseases and microorganisms are presented in a systematic and standardised manner using established reference resources to create database dictionaries. Detailed annotations regarding experimental designs, study cohorts, and the employed high- or low-throughput techniques are provided to cater to users' needs. Peryton offers several functionalities to enhance user experience and facilitate creative utilisation of the database. Users can query one or more microorganisms and/or diseases simultaneously. Advanced filtering options and direct text-based filtering enable precise refinement of search results and the execution of customised queries suited to different research inquiries. Peryton also provides interactive visualisations to effectively represent various aspects of its content, and users can download the results for local storage and further analysis. By serving as a valuable resource, Peryton enables researchers in the field of microbe-related diseases to generate new hypotheses and, importantly, facilitates the cross-validation of research findings. It is freely available online at www.microrna.gr/peryton. My main contribution to this project was the development of the database along with the back- and frontend.

Data Collection and Curation

Peryton is a meticulously curated database that focuses on experimentally supported microbe-disease associations. The database includes associations from studies covering neurodegenerative diseases, gastrointestinal disorders, cardiovascular diseases, and cancer. To ensure high quality, 2000 publication entries from PubMed database were manually curated from primary research articles, and only statistically significant associations with detailed experimental information were included. This comprehensive literature search and refinement process resulted in 314 publications that met the criteria for inclusion in the database. Peryton serves as a reliable resource for exploring microbe-disease associations in these specific disease categories.

Data Statistics

The curation process resulted in the identification of over 7,900 microbe-disease associations. These associations can be categorised into two groups: 3,777 entries (47.35%) comparing disease groups with healthy individuals, and 4,200 entries (52.65%) comparing different disease states. Peryton database covers 43 diseases and 1,396 microorganisms, including 23 cancer types, 10 gastrointestinal disorders, 7 cardiovascular diseases, and 3 neurodegenerative

disorders. The associations span across various taxonomic ranks, with a significant proportion (46%) being genus-related associations. Notably, 21.54% of the entries provide information at the species level or below. Each entry in the database includes detailed annotations such as bacterial abundance, study groups, experimental design, sample types, sequencing techniques used, Next-Generation Sequencing (NGS) sample accession numbers, and article metadata. It is worth mentioning that over 50% of the associations have a sample size of more than 50 individuals. The diseases and microorganisms in Peryton are presented in a standardised manner using MeSH [95] and NCBI Taxonomy [109] vocabularies respectively.

Database Implementation

Peryton is a relational database that follows the MVC (Model-View-Controller) architecture and is hosted on Apache HTTP server 2.4. The back-end of Peryton utilizes PostgreSQL server 11.8 (https://www.postgresql.org/) and the PHP framework Laravel 7.14 (https://laravel.com/) (PHP 7.2). On the front-end, the user interface is designed using Angular 9.1 (https://angular.io/) and incorporates the Angular Material UI library (https://material.angular.io/). The data in Peryton is stored in multiple tables within the PostgreSQL database, and Laravel handles the connection to these tables for data storage and retrieval. For data visualisation, Chart JS (https://www.chartjs.org/) is used for presenting database statistics, while Flourish (https://flourish.studio/) is employed for more complex visualisations. The database design schema revolves around the associations, emphasising the avoidance of data redundancy while staying flexible during querying (Figure 24).

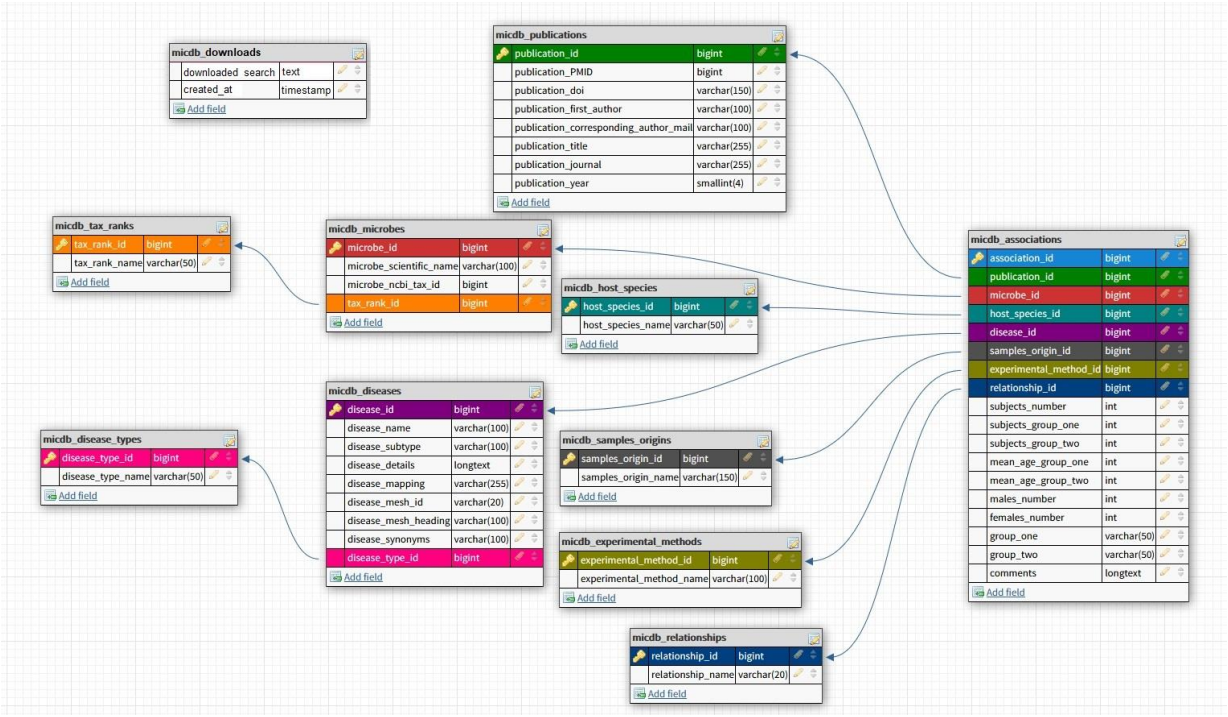

Figure 24. *The Peryton database schema (figure created for the needs of the current thesis)*

Interface and Functionality

We have developed a user-friendly interface for Peryton, which offers various functionalities to improve the user experience and allow creative utilisation of the database (Figure 24). Users have the option to query one or multiple microorganisms and/or diseases simultaneously. Advanced filtering options are available to refine search results based on taxonomic rank, experimental methodology, disease type, sample type, sample size, publication year, and relative abundance type (Figure 25A). Users can also choose to retrieve associations only in comparison to healthy individuals. The interface supports direct text-based filtering of results, enabling users to refine the information returned and perform customised queries based on their research needs. Peryton is integrated with the NCBI Taxonomy database [109], PubMed database, and MeSH terms [95], providing hyperlinks for direct access to relevant information on microorganisms, publications, and diseases, respectively (Figure 25B). Furthermore, we have compiled a list of common contaminants, following Eisenhofer et al. [110], which is integrated into the database. This feature allows users to identify whether a microorganism involved in an association is considered a common contaminant, requiring additional caution in interpreting and handling the results.

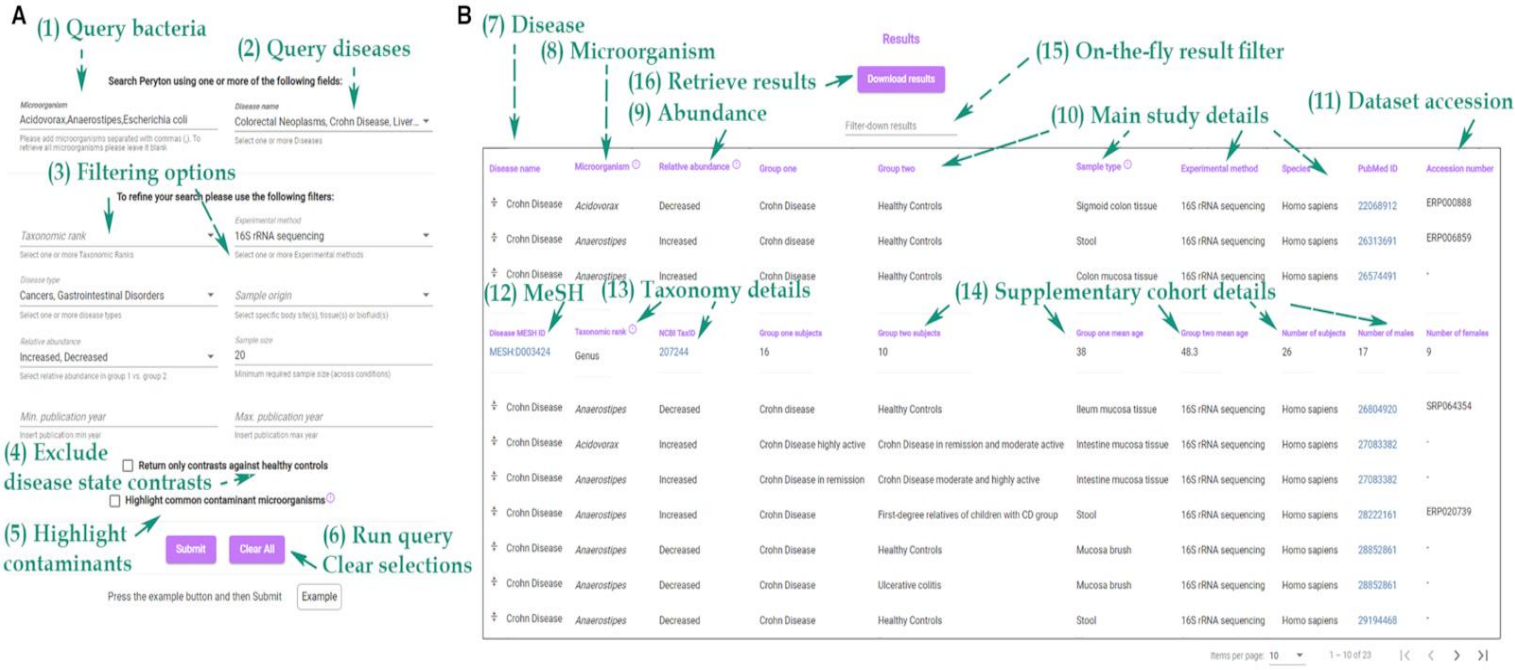

Figure 25. *The Peryton database interface* [108]

Peryton also offers advanced interactive visualisations to effectively explore its content and gain insights into microbe-disease relationships. Through network graphs, chord diagrams, and hierarchy diagrams, users can navigate the database and analyse information from relevant literature. The interactive graph network visualisation (Figure 26A) displays the strongest microbe-disease associations, with color-coded nodes representing diseases and microorganisms grouped by taxonomic rank while also providing relevant information and link navigation to relevant associations or external resources. The chord diagram (Figure 26B) allows users to observe and compare microbe-disease associations, with string widths reflecting the number of supporting entries. The hierarchy diagram (Figure 26C) is a multi-layered tool that provides information on all entries. Users can explore diseases and the number of entries within each taxonomic rank with a zoom-in functionality provided upon click, ultimately leading to detailed information on a specific disease, linked microorganisms, and the frequency of each association.

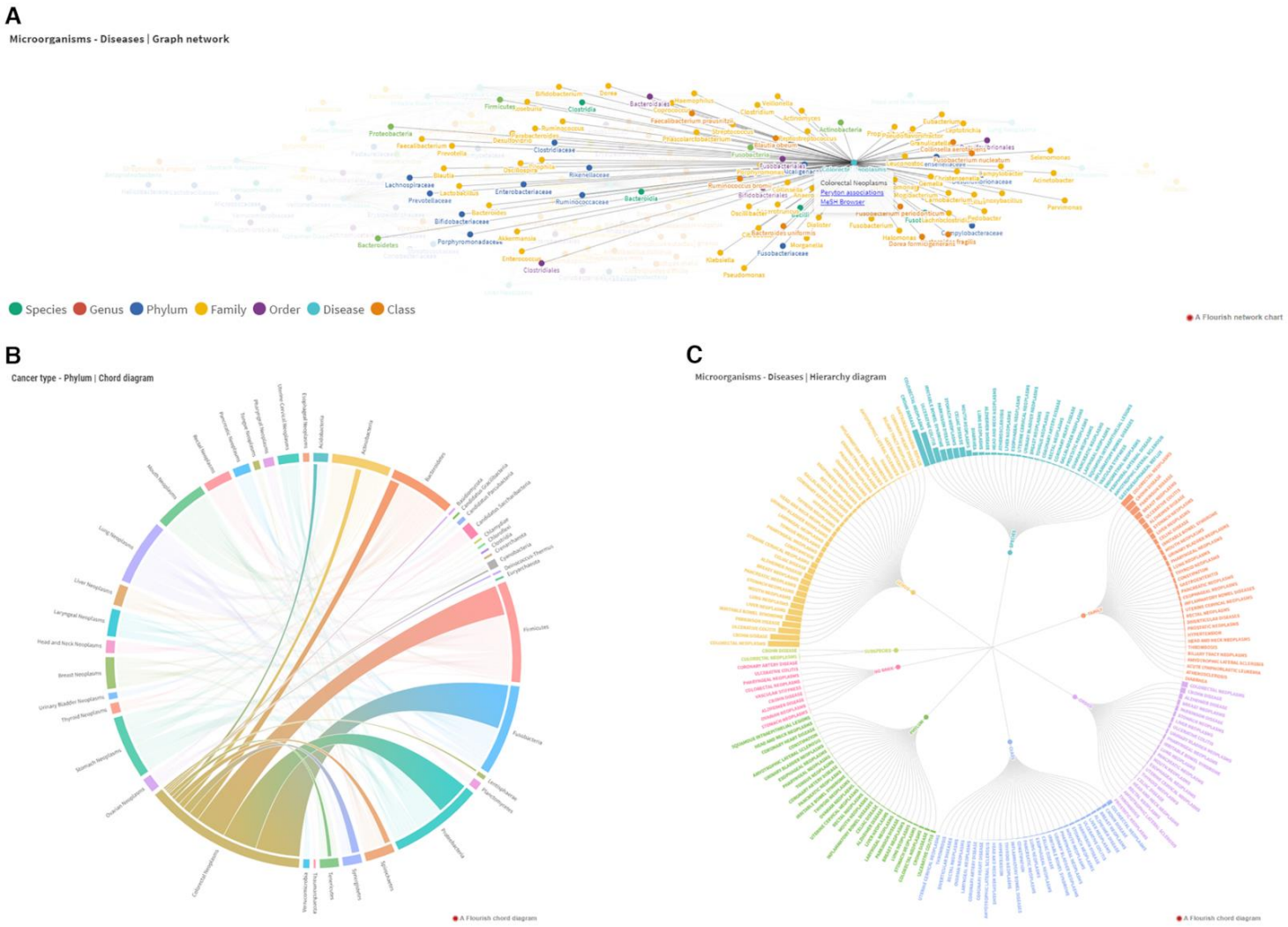

Figure 26. *The Peryton database visualisations* [108]

Finally, users have the option to download results from Peryton as tab-separated files, which include additional metadata such as disease and publication information. These files can be stored locally and used for further analysis. Basic database statistics, such as the top 15 diseases and top 15 microorganisms, are presented through bar plots on a dedicated page. A help page is available, providing screenshots and walkthroughs to guide users and ensure a smooth experience, particularly for first-time visitors. Additionally, there is a dedicated page where users can submit their own associations. They can choose between an advanced submission form with multiple association fields or a simple form requiring only the publication's DOI, PubMed ID, and a submitter comment. User-submitted associations will be considered by Peryton's curation team in future updates of the database content.

2.2.5. Agnodice

Overview

Agnodice is a meticulously curated database of bacterial sRNA-RNA interactions available at http://www.microrna.gr/agnodice. The initial version of Agnodice encompasses approximately 22,000 entries, providing comprehensive annotations at strain-level resolution. It encompasses around 390 small RNAs (sRNAs) and 6,630 target RNAs identified across 78 bacterial strains. The database exclusively includes experimentally supported interactions obtained through various methodologies, ranging from low yield to advanced high-throughput techniques. Agnodice offers a range of functionalities to enhance user experience and facilitate insightful utilisation. Each annotated sRNA-RNA interaction is accompanied by extensive metadata, catering to the needs of users. The database allows simultaneous querying of one or more sRNAs and/or target RNAs, with advanced filtering options available to refine search results and tailor queries for different research scenarios. Furthermore, Agnodice provides interactive visualisation tools specifically focused on the strongest sRNA-RNA interactions in model species such as *Escherichia coli*, *Salmonella enterica*, and *Pseudomonas aeruginosa*. The entire database content is freely accessible and can be directly downloaded for further analyses. Agnodice serves as a valuable resource, enabling microbiologists to generate novel hypotheses, identify/design sRNA-based drug targets, and explore the therapeutic potential of microbiomes from the perspective of small regulatory RNAs. Agnodice is currently under review process for publication. My contribution to this project was the development of the database along with the back- and frontend.

Data Collection and Curation

In order to gather a comprehensive collection of interactomic NGS datasets, we conducted extensive PubMed queries using specific keywords. This approach led to the identification of 24 datasets utilising various NGS methodologies for studying sRNA-RNA interactions. To further refine the collection and focus on experimentally confirmed interactions, we developed a text mining pipeline with full-text capability. Using a curated set of relevant scientific publications, we applied pre-processing and evaluated the full-text articles to identify sentences containing the desired associations. Through meticulous manual curation, we created a table with 33 fields to record interactions verified by low-throughput methods. The final collection underwent independent cross-checking and post-processing to ensure consistency and quality.

The criteria for recording an interaction entry included rigorous experimental support, detailed information on the experimental design and components, and statistical significance when applicable. Through this process, we derived approximately 1,000 interactions from low-yield methodologies, involving over 70 bacterial strains. For interactions derived from high-throughput experiments, custom-made pipelines were constructed using Python, incorporating supplementary files and annotation data from NCBI [109] and BioCyc [111] databases to populate the database fields.

Data Statistics

Through extensive manual curation of over 100 studies, a collection of approximately 22,000 sRNA-RNA interactions have been compiled in the Agnodice database. This database includes a total of 12,230 unique sRNA-RNA pairs involving 390 sRNAs and around 6,630 coding and non-coding RNAs. Among these entries, more than 1,000 interactions have been derived from high-confidence methodologies specifically designed to assess RNA-RNA binding sites. Agnodice also incorporates approximately 4,550 coding RNAs, which are annotated with 4,100 unique RNA products (i.e. proteins), as well as around 130 non-coding RNAs. These interactions have been obtained through 45 experimental methods, encompassing both high-throughput and low-throughput approaches. The database covers interactions mediated by three different RNA-binding proteins (RBPs): Hfq, CsrA, and ProQ, as well as interactions that do not have evidence of RBP involvement. Each annotated interaction in Agnodice is accompanied by additional information such as article metadata, microorganism names and TaxIDs (derived exclusively from the NCBI Taxonomy database [109]), bacterial lineage details (including phylum, genus, species, and strain names), information on the experimental methods employed, and more. Furthermore, Agnodice provides comprehensive details about the sRNAs, including synonym names, coordinates, sequence, upstream and downstream flanking genes, and strand information.

Database Implementation

Agnodice was built using the MVC architecture as a relational database and is being hosted on Apache HTTP server 2.4. The backend consists of a PostgreSQL server 11.8 (https://www.postgresql.org/) where *Agnodice's* data is stored in multiple tables featuring relational connections for optimal storing and querying. The PHP framework Laravel 8

(https://laravel.com/) (PHP 7.2) handles the backend logic including the connection to the PostgreSQL server for the storing and retrieval of the data. The frontend is designed as a one-page website using Angular 14 (https://angular.io/) and the Angular Material UI library (https://material.angular.io/). Finally, the database statistics are presented using the Chart JS (https://www.chartjs.org/) library, while Flourish (https://flourish.studio/) is utilised for the more complex visualisations provided. The database schema was designed to minimise duplication in data storage across the relational tables, but also focusing on the envisioned structure of the query results in order to optimise the table joining process and reduce the computational cost to produce it (Figure 27).

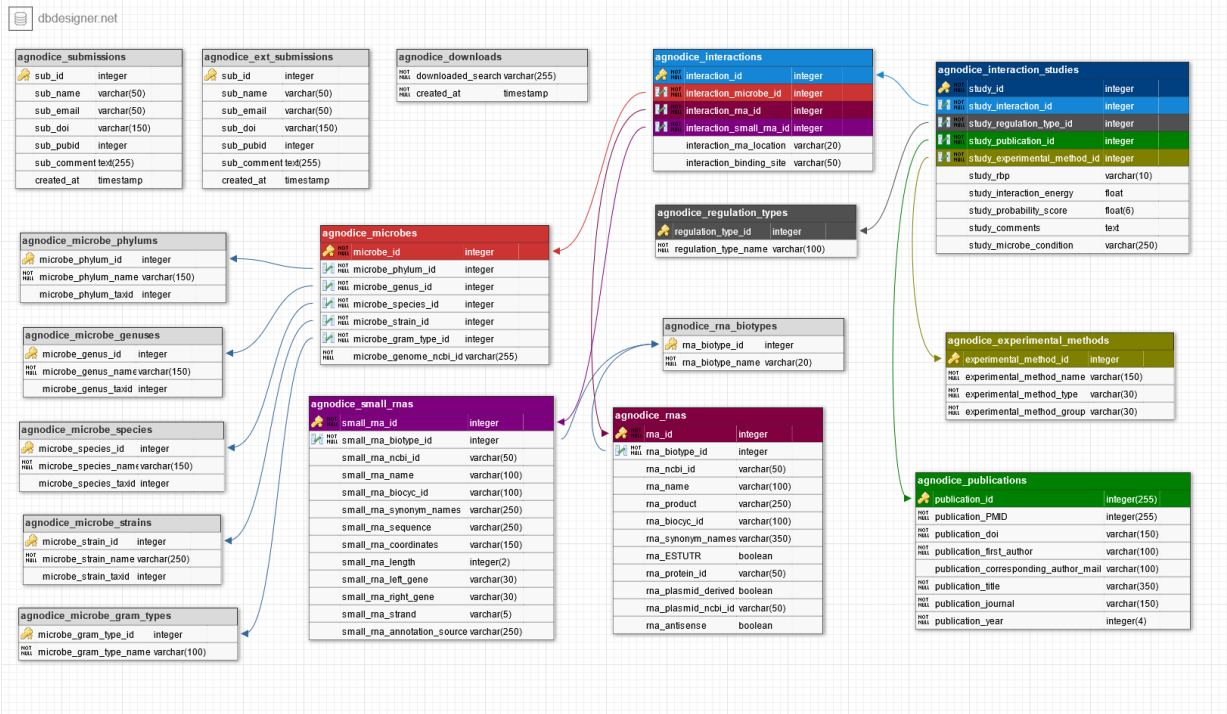

Figure 27. *The Agnodice database schema (figure created for the needs of the current thesis)*

Interface and Functionality

Agnodice is designed with a user-friendly interface to provide researchers with a comprehensive set of functionalities and access to an all-in-one resource for investigating sRNA-RNA interactions and addressing complex biological questions. The database offers support for queries using one or more sRNAs, genes, and microorganism names at different taxonomic ranks such as genus, species, and strain (Figure 28A). It also includes smart filtering options, such as "regulation type" (activation, repression, or unknown), "experimental method

name" and "publication years" allowing researchers to refine their search. Additionally, the filtering options include a checkbox to specifically include interactions derived from NGS-based techniques that directly assess sRNA-RNA binding events.

The results interface provides the interactions in a nested list format that collects interactions with common microorganism, sRNA and gene into one listing with an insight on variability of publications and experimental methods in each entry (Figure 28B). Upon collapse, each listing reveals all the different interactions included along with their specific information. Moreover, agnodice provides detailed information on the microorganisms, sRNAs and genes through pop-up windows. Finally, it strives to establish direct connections with various valuable resources, such as the NCBI Taxonomy reference organism indices [109], PubMed database for accessing relevant publications, and Peryton [108], a cutting-edge database of experimentally supported microbe-disease associations.

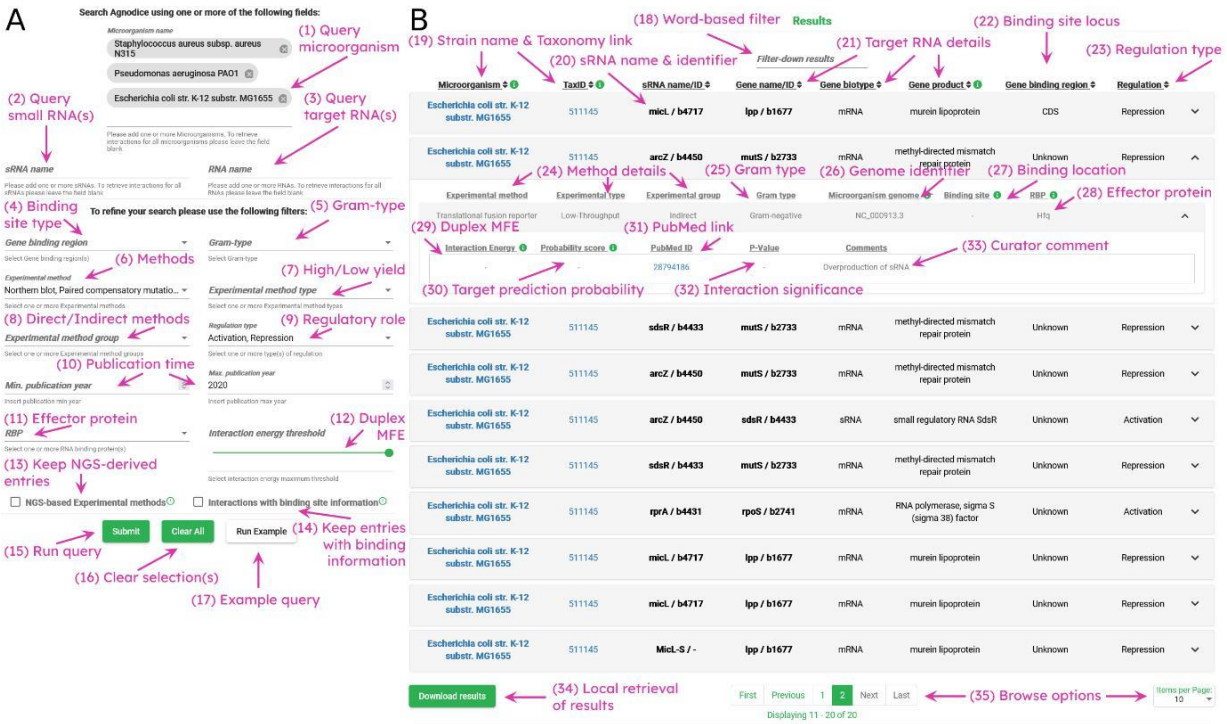

Figure 28. *The Agnodice query (**A**) and results (**B**) interface (figure created for the needs of the current thesis)*

A supplemental free-browse mode has been developed in Agnodice, allowing users to explore the entire collection of interactions without needing to perform a query or apply filters. A Visualisations page offers Chord diagrams that highlight the strongest sRNA-RNA interactions, specifically those that have been independently replicated in multiple studies, for

species like *E. coli*, *S. enterica*, and *P. aeruginosa*. The DB Statistics page provides the top 12 entries for key components of Agnodice, including microorganisms, sRNAs, genes, and experimental methods.

Users have the option to download all query results without any restrictions, along with additional metadata such as publication information and bacterial lineage details. This allows researchers to store Agnodice results locally and incorporate them into their own meta-analysis scenarios. A comprehensive Help page is available to guide users through the resource and assist them in performing tasks effortlessly. Additionally, researchers can submit their own interactions through a provided web form, which will be stored and manually reviewed by curators for the next major update of the database. This enables the scientific community to contribute and aid in the development of a centralised information hub for bacterial sRNA-RNA interactions, facilitating the initiation of new experiments.

3. Discussion and Conclusion

In an ever-expanding field of bioinformatics and an era where the advancements in technology facilitate the rapid growth of biological data accumulation, this dissertation focused on providing tools and resources for the identification and study of non-coding biomarkers in disease. The differences in the produced biological data based on the production medium as well as the vast number of bioinformatic tools and their variable degrees in ease of use frequently leave researchers unable to properly harness the power of big data analysis. The tools created aim to provide automated, streamlined, modular and robust analysis of high-throughput RNA and miRNA data while also introducing novel meticulous pre-processing of the raw data still in an automated way. The latter further increases the researcher's ability to include the vast amount of publicly available biological data in their studies with minimal effort and irrelevant of the amount of metadata provided with them. Furthermore, since many tools are created with a focus on specific data and/or organism analysis, a pipeline that provides multiple workflows utilising some of the best available tools, allows for critical choices for the analysis depending on the input data, the required results and even the available computational power.

The specialisation of studies in the present day on one hand provides further insight on specific topics, but on the other hand ends up spreading the overall scientific advancements of a field into an extensive number of publications, limiting one's ability to see the wider picture and combine the results without a considerable amount of effort. The resources developed as part of this dissertation focused on meticulously gathering this information and through manual curation providing the results of countless studies in a structured and uniform way. These resources range from miRNA expression data and experimentally supported biomarker functionality, to computationally predicted miRNA-RNA interactions and even microbial associations with disease. Providing the ability to query the collective result data of a wide amount of studies in a user-friendly way and give back structured, interconnected results is sure to help facilitate answering complex biological questions and forming research hypotheses.

Publications and Academic Activity

Scientific Publications

- **Alexiou A**, Zisis D, Kavakiotis I, Miliotis M, Koussounadis A, Karagkouni D, Hatzigeorgiou AG. *DIANA-mAP: Analyzing miRNA from raw NGS data to quantification.* Genes. 2020 Dec 30;12(1):46.
- Tastsoglou S, **Alexiou A**, Karagkouni D, Skoufos G, Zacharopoulou E, Hatzigeorgiou AG. *DIANA-microT 2023: including predicted targets of virally encoded miRNAs*. Nucleic Acids Research. 2023 Apr 24:gkad283.
- Kavakiotis I, **Alexiou A**, Tastsoglou S, Vlachos IS, Hatzigeorgiou AG. *DIANA-miTED: a microRNA tissue expression database*. Nucleic Acids Research. 2022 Jan 7;50(D1):D1055-61.
- Skoufos G, Kardaras FS, **Alexiou A**, Kavakiotis I, Lambropoulou A, Kotsira V, Tastsoglou S, Hatzigeorgiou AG. *Peryton: a manual collection of experimentally supported microbe-disease associations*. Nucleic acids research. 2021 Jan 8;49(D1):D1328-33.
- Tastsoglou S, Miliotis M, Kavakiotis I, **Alexiou A**, Gkotsi EC, Lambropoulou A, Lygnos V, Kotsira V, Maroulis V, Zisis D, Skoufos G. *Plasmir: A manual collection of circulating micrornas of prognostic and diagnostic value*. Cancers. 2021 Jul 22;13(15):3680.

Conference Participation

- **European Conference of Computational Biology (ECCB) 2018 - Athens, Greece - Sep 2018**:
  Oral presentation and Poster - "*Automated MicroRNA NGS Analysis Pipeline: From pre-process to quantification*"
- **Elixir All-hands meeting 2019 - Lisbon, Portugal - Jun 2019**:
  Poster presentation - "*DIANA-mAP: Analysing miRNA from raw NGS data to quantification*"

- **14th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB19) - Patra, Greece - Dec 2019**:
  Best Poster Presentation Award - "*DIANA-mAP: Analysing miRNA from raw NGS data to quantification*"
- **European Conference of Computational Biology (ECCB) 2022 - Barcelona, Spain - Sep 2022**:
  Poster presentation - "*Plasmir: A manual collection of circulating microrna of prognostic and diagnostic value*"
- **16th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB22) - Alexandroupoli, Greece - Oct 2022**:
  Poster Presentation - "*Agnodice: an expanded catalogue of experimentally supported bacterial sRNA-RNA interactions*"

References

[1] B. C. McKay, “Post-Transcriptional Regulation of DNA Damage-Responsive Gene Expression,” *Antioxid. Redox Signal.*, vol. 20, no. 4, pp. 640–654, Feb. 2014, doi: 10.1089/ars.2013.5523.

[2] S. Li and C. E. Mason, “The Pivotal Regulatory Landscape of RNA Modifications,” *Annu. Rev. Genomics Hum. Genet.*, vol. 15, no. 1, pp. 127–150, Aug. 2014, doi: 10.1146/annurev-genom-090413-025405.

[3] L. He and G. J. Hannon, “MicroRNAs: small RNAs with a big role in gene regulation,” *Nat. Rev. Genet.*, vol. 5, no. 7, pp. 522–531, Jul. 2004, doi: 10.1038/nrg1379.

[4] E. Van Rooij, A. L. Purcell, and A. A. Levin, “Developing MicroRNA Therapeutics,” *Circ. Res.*, vol. 110, no. 3, pp. 496–507, Feb. 2012, doi: 10.1161/CIRCRESAHA.111.247916.

[5] T. Kramps and K. Elbers, “Introduction to RNA Vaccines,” in *RNA Vaccines*, T. Kramps and K. Elbers, Eds., in Methods in Molecular Biology, vol. 1499. New York, NY: Springer New York, 2017, pp. 1–11. doi: 10.1007/978-1-4939-6481-9_1.

[6] M. J. Moore, “From Birth to Death: The Complex Lives of Eukaryotic mRNAs,” *Science*, vol. 309, no. 5740, pp. 1514–1518, Sep. 2005, doi: 10.1126/science.1111443.

[7] D. Söll and U. RajBhandary, Eds., *tRNA: structure, biosynthesis, and function*. Washington, D.C: ASM Press, 1995.

[8] R. Brimacombe and W. Stiege, “Structure and function of ribosomal RNA,” *Biochem. J.*, vol. 229, no. 1, pp. 1–17, Jul. 1985, doi: 10.1042/bj2290001.

[9] R. Reddy and H. Busch, “Small Nuclear RNAs: RNA Sequences, Structure, and Modifications,” in *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*, M. L. Birnstiel, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 1988, pp. 1–37. doi: 10.1007/978-3-642-73020-7_1.

[10] L.-A. MacFarlane and P. R. Murphy, “MicroRNA: Biogenesis, Function and Role in Cancer,” *Curr. Genomics*, vol. 11, no. 7, pp. 537–561, Nov. 2010, doi: 10.2174/138920210793175895.

[11] T. R. Mercer, M. E. Dinger, and J. S. Mattick, “Long non-coding RNAs: insights into functions,” *Nat. Rev. Genet.*, vol. 10, no. 3, pp. 155–159, Mar. 2009, doi: 10.1038/nrg2521.

[12] J.-H. Yoon, K. Abdelmohsen, and M. Gorospe, “Functional interactions among microRNAs and long noncoding RNAs,” *Semin. Cell Dev. Biol.*, vol. 34, pp. 9–14, Oct. 2014, doi: 10.1016/j.semcdb.2014.05.015.

[13] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin, “Computational identification of Drosophila microRNA genes,” *Genome Biol.*, vol. 4, no. 7, p. R42, 2003, doi: 10.1186/gb-2003-4-7-r42.

[14] A. K. Pradhan, L. Emdad, S. K. Das, D. Sarkar, and P. B. Fisher, “The Enigma of miRNA Regulation in Cancer,” in *Advances in Cancer Research*, Elsevier, 2017, pp. 25–52. doi: 10.1016/bs.acr.2017.06.001.

[15] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, “Most mammalian mRNAs are conserved targets of microRNAs,” *Genome Res.*, vol. 19, no. 1, pp. 92–105, Jan. 2009, doi: 10.1101/gr.082701.108.

[16] I. S. Vlachos, G. Georgakilas, S. Tastsoglou, M. D. Paraskevopoulou, D. Karagkouni, and A. G. Hatzigeorgiou, “Computational Challenges and -omics Approaches for the Identification of microRNAs and Targets,” in *Essentials of Noncoding RNA in Neuroscience*, Elsevier, 2017, pp. 39–59. doi: 10.1016/B978-0-12-804402-5.00003-0.

[17] D. Sayed and M. Abdellatif, “MicroRNAs in Development and Disease,” *Physiol. Rev.*, vol. 91, no. 3, pp. 827–887, Jul. 2011, doi: 10.1152/physrev.00006.2010.

[18] M. D. Jansson and A. H. Lund, “MicroRNA and cancer,” *Mol. Oncol.*, vol. 6, no. 6, pp.

590–610, Dec. 2012, doi: 10.1016/j.molonc.2012.09.006.
[19] L. Ma, “MicroRNA and Metastasis,” in *Advances in Cancer Research*, Elsevier, 2016, pp. 165–207. doi: 10.1016/bs.acr.2016.07.004.
[20] C. E. Condrat *et al.*, “miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis,” *Cells*, vol. 9, no. 2, p. 276, Jan. 2020, doi: 10.3390/cells9020276.
[21] M.-C. King, J. H. Marks, and J. B. Mandell, “Breast and Ovarian Cancer Risks Due to Inherited Mutations in *BRCA1* and *BRCA2*,” *Science*, vol. 302, no. 5645, pp. 643–646, Oct. 2003, doi: 10.1126/science.1088759.
[22] W. J. Catalona *et al.*, “Measurement of Prostate-Specific Antigen in Serum as a Screening Test for Prostate Cancer,” *N. Engl. J. Med.*, vol. 324, no. 17, pp. 1156–1161, Apr. 1991, doi: 10.1056/NEJM199104253241702.
[23] K. Herholz and K. Ebmeier, “Clinical amyloid imaging in Alzheimer’s disease,” *Lancet Neurol.*, vol. 10, no. 7, pp. 667–670, Jul. 2011, doi: 10.1016/S1474-4422(11)70123-5.
[24] N. Ford, G. Meintjes, M. Vitoria, G. Greene, and T. Chiller, “The evolving role of CD4 cell counts in HIV care,” *Curr. Opin. HIV AIDS*, vol. 12, no. 2, pp. 123–128, Mar. 2017, doi: 10.1097/COH.0000000000000348.
[25] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proc. Natl. Acad. Sci.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.
[26] E. Y. Chan, “Advances in sequencing technology,” *Mutat. Res. Mol. Mech. Mutagen.*, vol. 573, no. 1–2, pp. 13–40, Jun. 2005, doi: 10.1016/j.mrfmmm.2005.01.004.
[27] E. R. Mardis, “Next-Generation Sequencing Platforms,” *Annu. Rev. Anal. Chem.*, vol. 6, no. 1, pp. 287–303, Jun. 2013, doi: 10.1146/annurev-anchem-062012-092628.
[28] D. Xu, J. Caballero, G. Glusman, and Q. Tian, “Diagnostic perspectives in the epoch of next-generation sequencing,” in *Next-Generation Sequencing & Molecular Diagnostics*, Unitec House, 2 Albert Place, London N3 1QB, UK: Future Medicine Ltd, 2013, pp. 98–111. doi: 10.2217/ebo.12.329.
[29] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, “The Next-Generation Sequencing Revolution and Its Impact on Genomics,” *Cell*, vol. 155, no. 1, pp. 27–38, Sep. 2013, doi: 10.1016/j.cell.2013.09.006.
[30] A. Esposito, C. Colantuono, V. Ruggieri, and M. L. Chiusano, “Bioinformatics for agriculture in the Next-Generation sequencing era,” *Chem. Biol. Technol. Agric.*, vol. 3, no. 1, p. 9, Dec. 2016, doi: 10.1186/s40538-016-0054-8.
[31] X. Wang, *Next-generation sequencing data analysis*. Boca Raton: Taylor & Francis, 2016.
[32] S. Aryal, “Next-Generation Sequencing (NGS)- Definition, Types,” *Next-Generation Sequencing (NGS)- Definition, Types*. https://microbenotes.com/next-generation-sequencing-ngs/
[33] K. R. Kumar, M. J. Cowley, and R. L. Davis, “Next-Generation Sequencing and Emerging Technologies,” *Semin. Thromb. Hemost.*, vol. 45, no. 07, pp. 661–673, Oct. 2019, doi: 10.1055/s-0039-1688446.
[34] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.
[35] K. R. Kukurba and S. B. Montgomery, “RNA Sequencing and Analysis,” *Cold Spring Harb. Protoc.*, vol. 2015, no. 11, p. pdb.top084970, Nov. 2015, doi: 10.1101/pdb.top084970.
[36] K. P. McCormick, M. R. Willmann, and B. C. Meyers, “Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing

experiments," *Silence*, vol. 2, no. 1, p. 2, Dec. 2011, doi: 10.1186/1758-907X-2-2.
[37] S. Benesova, M. Kubista, and L. Valihrach, "Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis," *Diagnostics*, vol. 11, no. 6, p. 964, May 2021, doi: 10.3390/diagnostics11060964.
[38] C. Trapnell *et al.*, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nat. Protoc.*, vol. 7, no. 3, Art. no. 3, Mar. 2012, doi: 10.1038/nprot.2012.016.
[39] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," *Nat. Protoc.*, vol. 11, no. 9, Art. no. 9, Sep. 2016, doi: 10.1038/nprot.2016.095.
[40] X. Wu *et al.*, "sRNAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline," *Nucleic Acids Res.*, vol. 45, no. 21, pp. 12140–12151, Dec. 2017, doi: 10.1093/nar/gkx999.
[41] S. Zhao *et al.*, "QuickMIRSeq: a pipeline for quick and accurate quantification of both known miRNAs and isomiRs by jointly processing multiple samples from microRNA sequencing," *BMC Bioinformatics*, vol. 18, no. 1, p. 180, Mar. 2017, doi: 10.1186/s12859-017-1601-4.
[42] T. Desvignes, P. Batzel, J. Sydes, B. F. Eames, and J. H. Postlethwait, "miRNA analysis with Prost! reveals evolutionary conservation of organ-enriched expression and post-transcriptional modifications in three-spined stickleback and zebrafish," *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41598-019-40361-8.
[43] X. Zhong, A. Pla, and S. Rayner, "Jasmine: a Java pipeline for isomiR characterization in miRNA-Seq data," *Bioinformatics*, vol. 36, no. 6, pp. 1933–1936, Mar. 2020, doi: 10.1093/bioinformatics/btz806.
[44] A. Rueda *et al.*, "sRNAtoolbox: an integrated collection of small RNA research tools," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W467–W473, Jul. 2015, doi: 10.1093/nar/gkv555.
[45] J. Wu *et al.*, "mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing," *RNA Biol.*, vol. 10, no. 7, pp. 1087–1092, Jul. 2013, doi: 10.4161/rna.25193.
[46] Z. Sun *et al.*, "CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data," *BMC Genomics*, vol. 15, no. 1, p. 423, Jun. 2014, doi: 10.1186/1471-2164-15-423.
[47] Y. Lu, A. S. Baras, and M. K. Halushka, "miRge 2.0 for comprehensive analysis of microRNA sequencing data," *BMC Bioinformatics*, vol. 19, no. 1, p. 275, Jul. 2018, doi: 10.1186/s12859-018-2287-y.
[48] E. Andrés-León, R. Núñez-Torres, and A. M. Rojas, "miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis," *Sci. Rep.*, vol. 6, no. 1, Art. no. 1, May 2016, doi: 10.1038/srep25749.
[49] M. P. A. Davis, S. van Dongen, C. Abreu-Goodger, N. Bartonicek, and A. J. Enright, "Kraken: A set of tools for quality control and analysis of high-throughput sequence data," *Methods*, vol. 63, no. 1, pp. 41–49, Sep. 2013, doi: 10.1016/j.ymeth.2013.06.027.
[50] B. Panwar, G. S. Omenn, and Y. Guan, "miRmine: a database of human miRNA expression profiles," *Bioinformatics*, vol. 33, no. 10, pp. 1554–1560, May 2017, doi: 10.1093/bioinformatics/btx019.
[51] J. Gong *et al.*, "Comprehensive analysis of human small RNA sequencing data provides insights into expression profiles and miRNA editing," *RNA Biol.*, vol. 11, no. 11, pp. 1375–1385, Nov. 2014, doi: 10.1080/15476286.2014.996465.
[52] P. P. Kuksa, A. Amlie-Wolf, Ž. Katanić, O. Valladares, L.-S. Wang, and Y. Y. Leung, "DASHR 2.0: integrated database of human small non-coding RNA genes and mature products," *Bioinformatics*, vol. 35, no. 6, pp. 1033–1039, Mar. 2019, doi:

10.1093/bioinformatics/bty709.
[53] I.-F. Chung *et al.*, "YM500v3: a database for small RNA sequencing in human cancer research," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D925–D931, Jan. 2017, doi: 10.1093/nar/gkw1084.
[54] R.-U. Rahman *et al.*, "SEAweb: the small RNA Expression Atlas web application," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D204–D219, Jan. 2020, doi: 10.1093/nar/gkz869.
[55] F. Xie *et al.*, "deepBase v3.0: expression atlas and interactive analysis of ncRNAs from thousands of deep-sequencing data," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D877–D883, Jan. 2021, doi: 10.1093/nar/gkaa1039.
[56] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Współczesna Onkol.*, vol. 1A, pp. 68–77, 2015, doi: 10.5114/wo.2014.47136.
[57] J. Zhang *et al.*, "The International Cancer Genome Consortium Data Portal," *Nat. Biotechnol.*, vol. 37, no. 4, pp. 367–369, Apr. 2019, doi: 10.1038/s41587-019-0055-9.
[58] E. Clough and T. Barrett, "The Gene Expression Omnibus Database," in *Statistical Genomics: Methods and Protocols*, E. Mathé and S. Davis, Eds., in Methods in Molecular Biology. New York, NY: Springer, 2016, pp. 93–110. doi: 10.1007/978-1-4939-3578-9_5.
[59] R. Leinonen, H. Sugawara, M. Shumway, and on behalf of the I. N. S. D. Collaboration, "The Sequence Read Archive," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D19–D21, Jan. 2011, doi: 10.1093/nar/gkq1019.
[60] F. Russo *et al.*, "miRandola 2017: a curated knowledge base of non-invasive biomarkers," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D354–D359, Jan. 2018, doi: 10.1093/nar/gkx854.
[61] Z. Huang *et al.*, "HMDD v3.0: a database for experimentally supported human microRNA–disease associations," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1013–D1017, Jan. 2019, doi: 10.1093/nar/gky1010.
[62] J.-R. Li, C.-Y. Tong, T.-J. Sung, T.-Y. Kang, X. J. Zhou, and C.-C. Liu, "CMEP: a database for circulating microRNA expression profiling," *Bioinformatics*, vol. 35, no. 17, pp. 3127–3132, Sep. 2019, doi: 10.1093/bioinformatics/btz042.
[63] A. Alexiou *et al.*, "DIANA-mAP: Analyzing miRNA from Raw NGS Data to Quantification," *Genes*, vol. 12, no. 1, p. 46, Dec. 2020, doi: 10.3390/genes12010046.
[64] J. Tsuji and Z. Weng, "DNApi: A De Novo Adapter Prediction Algorithm for Small RNA Sequencing Data," *PLOS ONE*, vol. 11, no. 10, p. e0164228, Oct. 2016, doi: 10.1371/journal.pone.0164228.
[65] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, Mar. 2009, doi: 10.1186/gb-2009-10-3-r25.
[66] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades," *Nucleic Acids Res.*, vol. 40, no. 1, pp. 37–52, Jan. 2012, doi: 10.1093/nar/gkr688.
[67] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D152–D157, Jan. 2011, doi: 10.1093/nar/gkq1027.
[68] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
[69] T. E. P. Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project," *Science*, vol. 306, no. 5696, pp. 636–640, Oct. 2004, doi: 10.1126/science.1105136.
[70] "FastQC: a quality control tool for high throughput sequence data." 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[71] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, Art. no. 1, May 2011, doi: 10.14806/ej.17.1.200.
[72] C. Camps *et al.*, “Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia,” *Mol. Cancer*, vol. 13, no. 1, p. 28, Feb. 2014, doi: 10.1186/1476-4598-13-28.
[73] D. D. Jima *et al.*, “Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs,” *Blood*, vol. 116, no. 23, pp. e118–e127, Dec. 2010, doi: 10.1182/blood-2010-05-285403.
[74] I. S. Vlachos *et al.*, “DIANA-mirExTra v2.0: Uncovering microRNAs and transcription factors with crucial roles in NGS expression data,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W128–W134, Jul. 2016, doi: 10.1093/nar/gkw455.
[75] J. E. Handzlik, S. Tastsoglou, I. S. Vlachos, and A. G. Hatzigeorgiou, “Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data,” *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41598-020-57495-9.
[76] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nat. Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.
[77] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
[78] B. Li and C. N. Dewey, “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome,” *BMC Bioinformatics*, vol. 12, no. 1, p. 323, Dec. 2011, doi: 10.1186/1471-2105-12-323.
[79] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.
[80] Krueger, Felix, “Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data,” *Babraham Inst.*, 2015.
[81] P. Danecek *et al.*, “Twelve years of SAMtools and BCFtools,” *GigaScience*, vol. 10, no. 2, p. giab008, Jan. 2021, doi: 10.1093/gigascience/giab008.
[82] Y. Liao, G. K. Smyth, and W. Shi, “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features,” *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014, doi: 10.1093/bioinformatics/btt656.
[83] Broad Institute, “Picard toolkit,” *Broad Inst. GitHub Repos.*, 2019, [Online]. Available: https://broadinstitute.github.io/picard/
[84] L. Wang, S. Wang, and W. Li, “RSeQC: quality control of RNA-seq experiments,” *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, Aug. 2012, doi: 10.1093/bioinformatics/bts356.
[85] I. Kavakiotis, A. Alexiou, S. Tastsoglou, I. S. Vlachos, and A. G. Hatzigeorgiou, “DIANA-miTED: a microRNA tissue expression database,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1055–D1061, Jan. 2022, doi: 10.1093/nar/gkab733.
[86] S. Tarallo *et al.*, “Altered Fecal Small RNA Profiles in Colorectal Cancer Reflect Gut Microbiome Composition in Stool Samples,” *mSystems*, vol. 4, no. 5, pp. e00289-19, Oct. 2019, doi: 10.1128/mSystems.00289-19.
[87] M. D. Paraskevopoulou *et al.*, “DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows,” *Nucleic Acids Res.*, vol. 41, no. W1, pp. W169–W173, Jul. 2013, doi: 10.1093/nar/gkt393.
[88] D. Karagkouni *et al.*, “DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D239–

D245, Jan. 2018, doi: 10.1093/nar/gkx1141.
[89] D. Karagkouni *et al.*, "DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts," *Nucleic Acids Res.*, p. gkz1036, Nov. 2019, doi: 10.1093/nar/gkz1036.
[90] I. S. Vlachos *et al.*, "DIANA-miRPath v3.0: deciphering microRNA function with experimental support," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W460–W466, Jul. 2015, doi: 10.1093/nar/gkv403.
[91] S. Tastsoglou *et al.*, "PlasmiR: A Manual Collection of Circulating microRNAs of Prognostic and Diagnostic Value," *Cancers*, vol. 13, no. 15, p. 3680, Jul. 2021, doi: 10.3390/cancers13153680.
[92] The RNAcentral Consortium *et al.*, "RNAcentral: a hub of information for non-coding RNA sequences," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D221–D229, Jan. 2019, doi: 10.1093/nar/gky1034.
[93] B. Fromm *et al.*, "MirGeneDB 2.0: the metazoan microRNA complement," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D132–D141, Jan. 2020, doi: 10.1093/nar/gkz885.
[94] A. P. Davis *et al.*, "Comparative Toxicogenomics Database (CTD): update 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1138–D1143, Jan. 2021, doi: 10.1093/nar/gkaa891.
[95] C. E. Lipscomb, "Medical Subject Headings (MeSH)," *Bull. Med. Libr. Assoc.*, vol. 88, no. 3, pp. 265–266, Jul. 2000.
[96] S. M. Bello *et al.*, "Augmenting the disease ontology improves and unifies disease annotations across species," *Dis. Model. Mech.*, p. dmm.032839, Jan. 2018, doi: 10.1242/dmm.032839.
[97] J. S. Amberger and A. Hamosh, "Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes," *Curr. Protoc. Bioinforma.*, vol. 58, no. 1, Jun. 2017, doi: 10.1002/cpbi.27.
[98] M. Reczko, M. Maragkakis, P. Alexiou, I. Grosse, and A. G. Hatzigeorgiou, "Functional microRNA targets in protein coding sequences," *Bioinformatics*, vol. 28, no. 6, pp. 771–776, Mar. 2012, doi: 10.1093/bioinformatics/bts043.
[99] S. Tastsoglou, A. Alexiou, D. Karagkouni, G. Skoufos, E. Zacharopoulou, and A. G. Hatzigeorgiou, "DIANA-microT 2023: including predicted targets of virally encoded miRNAs," *Nucleic Acids Res.*, p. gkad283, Apr. 2023, doi: 10.1093/nar/gkad283.
[100] F. Cunningham *et al.*, "Ensembl 2022," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D988–D995, Jan. 2022, doi: 10.1093/nar/gkab1049.
[101] B. T. Lee *et al.*, "The UCSC Genome Browser database: 2022 update," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1115–D1122, Jan. 2022, doi: 10.1093/nar/gkab959.
[102] S. T. Sherry, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/nar/29.1.308.
[103] M. J. Landrum *et al.*, "ClinVar: improvements to accessing data," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D835–D844, Jan. 2020, doi: 10.1093/nar/gkz972.
[104] Y. Ding, C. Y. Chan, and C. E. Lawrence, "Sfold web server for statistical folding and rational design of nucleic acids," *Nucleic Acids Res.*, vol. 32, no. Web Server, pp. W135–W141, Jul. 2004, doi: 10.1093/nar/gkh449.
[105] S. E. McGeary *et al.*, "The biochemical basis of microRNA targeting efficacy," *Science*, vol. 366, no. 6472, p. eaav1741, Dec. 2019, doi: 10.1126/science.aav1741.
[106] The GTEx Consortium *et al.*, "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, May 2015, doi: 10.1126/science.1262110.
[107] J. F. Söllner *et al.*, "An RNA-Seq atlas of gene expression in mouse and rat normal tissues," *Sci. Data*, vol. 4, no. 1, p. 170185, Dec. 2017, doi: 10.1038/sdata.2017.185.

[108] G. Skoufos *et al.*, “Peryton: a manual collection of experimentally supported microbe-disease associations,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1328–D1333, Jan. 2021, doi: 10.1093/nar/gkaa902.
[109] S. Federhen, “The NCBI Taxonomy database,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D136-143, Jan. 2012, doi: 10.1093/nar/gkr1178.
[110] R. Eisenhofer, J. J. Minich, C. Marotz, A. Cooper, R. Knight, and L. S. Weyrich, “Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations,” *Trends Microbiol.*, vol. 27, no. 2, pp. 105–117, Feb. 2019, doi: 10.1016/j.tim.2018.11.003.
[111] P. D. Karp *et al.*, “The BioCyc collection of microbial genomes and metabolic pathways,” *Brief. Bioinform.*, vol. 20, no. 4, pp. 1085–1093, Jul. 2019, doi: 10.1093/bib/bbx085.