UNIVERSIDAD DE GRANADA

DOCTORAL THESIS

Programa de Doctorado en Biomedicina

Development of computational tools for the detection of circulating small RNA biomarkers in cancer patients

---

Ernesto L. Aparicio Puerta

Supervisors:

Michael Hackenberg

Juan Antonio Marchal Corrales

Granada, December 2021

Departamento de Genética

FACULTAD DE CIENCIAS

Cátedra "Doctores Galera y Requena de Investigación en Células Madre Cancerígenas"

Editor: Universidad de Granada. Tesis Doctorales
Autor: Ernesto Luis Aparicio Puerta
ISBN: 978-84-1117-236-3
URI: http://hdl.handle.net/10481/72885

This doctoral thesis was carried out at the Department of Genetics, University of Granada in collaboration with the Differentiation, Regeneration and Cancer group from the IBIMER institute and ibs.Granada. The thesis was funded by the following programs, projects and institutions:

- i-PFIS PhD programme ("Contratos i-PFIS: doctorados IIS-empresa en Ciencias y Tecnologías de la salud de la convocatoria 2016 de la Acción Estratégica en Salud"), Instituto de Salud Carlos III (Ministerio de Economía y Competitividad) [IFI16/00041]

- International research stay grant ("Ayudas para estancias formativas 2017"), Consejería de Salud, Junta de Andalucıa [EF-0484-2017]

- Research mobility grant ("Movilidad de Personal Investigador contratado en el marco de las AES, 2019"), Instituto de Salud Carlos III (Ministerio de Economía y Competitividad) [MV19/00058]

- Fulbright Program, Predoctoral Research 2020 Call, Bureau of Educational and Cultural Affairs of the U.S. Department of State, United States of America

- Research project [PIE16/00045] ("Proyectos integrados de excelencia de los institutos de investigación sanitaria 2016"), Instituto de Salud Carlos III (Ministerio de Economía y Competitividad)

- Research project [PI-0533-2014] ("Desarrollo de un sistema de nanodiagnóstico basado en miRNAs/exosomas caracterısticos de células madre cancerígenas con valor pronóstico y predictivo en pacientes con melanoma maligno"), Fundación progreso y Salud /Consejería de Salud, Junta de Andalucía)

A Ana

A mi familia

A mis amigos

"Science is not about building a body of known 'facts'. It is a method for asking awkward questions and subjecting them to a reality-check, thus avoiding the human tendency to believe whatever makes us feel good."

Terry Pratchett, The Science Of Discworld

Glossary

back-end	In web development, this is the term used to refer to the server processing part of the software (data access, calculations, etc)
CDS	Coding sequence. Fragment of a gene sequence that codes for a protein
cfDNA	Cell-free DNA
circRNA	Circular RNA
CTCs	Circulating Tumor Cells
ctDNA	Circulating tumor DNA
Django	Web development framework based on Python
dPCR	Digital PCR
EVs	Extracellular vesicles
FDA	United States Food and Drug Administration
FDR	False discovery rate
front-end	In web development, this is the term used to refer to web browser processing part of the software (layout, visualization, etc)
GWAS	Genome wide association study. An observational study to find genomic variants associated with traits such as diseases
lncRNA	Long non-coding RNA
melanoma	Type of skin cancer that originates from melanocytes, the cells that produce pigmentation in hair or skin

miRBase	Primary database of miRNA sequences and annotations
miRGeneDB	Database of manually curated microRNA genes
miRNA	Endogenous small ncRNA molecules (∼22 nucleotides) that mediate translational repression of target mRNAs
mRNA	messenger RNA
MySQL	Open-source relational database management system based on the SQL language
ncRNA	Non-coding RNA
NGS	Next Generation Sequencing
NIH	National Institutes of Health. Federal agency of the United States in charge of health research
NTA	Non-templated addition. Nucleotides post-transcriptionally added to a miRNA sequence by nucleotidyl transferases
PCR	Polymerase Chain Reaction
RC	Read count
RPM	Reads Per Million
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription polymerase chain reaction
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive. NIH's primary archive of high-throughput sequencing data
TNM	Staging system of malignant tumors that describes the size and spread of cancer

TORQUE	Resource and queue manager that allows job scheduling and control
tRNA	Transfer RNA
UMI	Unique Molecular Identifier. Molecular barcodes used to tag DNA fragments in NGS sequencing
UR	Unique Reads
UTR	Untranslated region. In a strand of mRNA, sequences found at each end of the molecule that surround the coding sequence and that are not translated into protein
WGS	Whole genome sequencing
WHO	World Health Organization

Abstract

Cancer is the leading cause of premature death (0-69) in Spain and in the vast majority of countries in the European Union. Developing countries are less affected but numbers are expected to rise over the next few decades as a consequence of population aging and lifestyle changes. Besides the human losses, cancer is responsible for large economic impact on healthcare systems and companies. It will come as no surprise to learn that many pharmaceutical companies and public funding agencies target cancer as their most funded research topic.

Cancer is a complex set of diseases that are normally labeled using the tissue and/or cell type from which they originate. Each type is associated with very different prognosis and treatments but in most cases, patients would benefit from early detection, diagnosis and treatment since tumors tend to become more aggressive with time. In the United States, breast and colon cancer mortality rates have dropped $\sim$40% in the last 25 years in part as a consequence of achieving earlier diagnoses on average.

Traditionally, cancer was diagnosed using samples resected from tumor tissue, a procedure known as biopsy. Biopsies, however, need to be performed and assessed by specialized personnel, making it impractical to repeat and almost impossible to use as screening strategy. In this context, liquid biopsy, a testing approach that consists in sampling and exploration of blood or other bodily fluids as a surrogate for regular biopsies, has emerged as a viable alternative. Many different biological materials can be explored from blood samples and several studies have already reported associations between the abundance of certain biomolecules and the presence of tumors, cancer stage or prognosis. Particularly, there has been great in-

terest in miRNAs because they are well-characterized molecules, abundant in bodily fluids and relatively stable.

A common approach to propose new circulating miRNAs as disease biomarkers is to rely on Next Generation Sequencing data from blood, serum, plasma or extracellular vesicles. In this thesis, I developed and updated tools that can be used to: 1) process miRNA-seq experiments and obtain miRNA and isomiR profiles as well as tRNA properties; and 2) to perform quality control of miRNA-seq experiments by comparison to over 30,000 publicly available samples uniformly processed using *sRNAbench*. Finally, liquid biopsy miRNA-seq samples were compiled in a manually curated database to increase reproducibility and reusability of public datasets.

sRNAbench is one of the most used web-servers for analysis of miRNA-seq experiments, with almost 50 thousand jobs launched since the latest publication in 2019. Among other things, *sRNAbench* enables miRNA and isomiR profiling, mapping to the genome and usage of several indexes. In 2019 we updated this service to include new library preparation protocols, species and miRNA reference annotations. Since the original publication of *sRNAbench*, the amount of miRNA-seq studies in *GEO* has more than triplicated. In consequence, we adapted this service to allow fast processing of multiple samples with the same protocol and automated download from *SRA*. The differential expression tool, sRNAde, was also improved to include more methods and an interactive experience.

Although liquid biopsy miRNA-seq studies have been on the rise for some time already, articles and publicly available data are not always consistently reported if at all. As a solution, we took the initiative to manually curate and uniformly process 31 *SRA* studies containing over 1000

miRNA-seq samples. We organized this dataset into a publicly available database, *liqDB*, which can be queried to retrieve any subset of samples and to generate or test hypothesis. Insights gained from this process were very important to determine what we thought constituted a miRNA-seq study of good quality.

Quality control of miRNA-seq experiments is a frequently overlooked matter. Even if miRNA-seq is extensively used, most quality control metrics differ very little from those provided in regular RNA-seq studies. Taking advantage from the automated processing and quality control pipeline developed to populate *liqDB*, we analyzed a corpus of over 36,000 miRNA-seq samples. Using this corpus and our experience, we proposed 27 quality features that are displayed in the context of the reference corpus by means of percentiles. Users can also subset the reference to compare their samples to more tailored data that resembles their experiment. One of the predefined subsets corresponds to bodily fluid experiments so liquid biopsy samples can be assessed by comparison to those only. To our knowledge, *mirnaQC* is the first bioinformatics tool especially conceived for miRNA-seq quality control and the only NGS QC tool that relies on a background set of samples for relative assessment.

Finally, to proof the usefulness of the methods presented here, we applied them to a set of serum samples from melanoma patients, a type of skin cancer that originates in melanocytes, the cells that produce pigmentation. Potential melanoma sufferers could really benefit from early detection strategies since survival rates display staggering differences between stages: 5-year survival rate in the United States is 99% for localized disease (Stages 0, I and II), 66% in the case of regional spread and only 27% for patients

with distant spread.

In our exploration of the dataset using differential expression analysis we found several miRNAs that were either significantly more abundant in healthy controls or in patients of melanoma. We also found miRNAs that displayed a stark difference between controls and patients in an early stage, which indicates great potential as early detection biomarkers. Finally, we found differences in the uridylation levels of several miRNAs that correlated with disease progression.

Key words: miRNA, sequencing, liquid biopsy, cancer, melanoma, early detection, metastasis

Resumen

El cáncer es la principal causa de muerte prematura (0-69) en España y en la mayoría de países de la Unión Europea. Los países en desarrollo se encuentran menos afectados, pero se espera que sus cifras aumenten en las próximas décadas como consecuencia del envejecimiento de la población y los cambios en el estilo de vida. Además de las pérdidas humanas, el cáncer es responsable de un gran impacto económico en los sistemas de salud y en las empresas. Por lo tanto, no es de sorprender que muchas farmacéuticas y agencias de investigación hayan hecho del cáncer la enfermedad cuya investigación mejor se financia.

El cáncer es en realidad un conjunto heterogéneo de enfermedades que normalmente se denominan usando el tejido a partir del cual se originan. El pronóstico varía mucho en función del tipo de cáncer pero en la inmensa mayoría de casos, los pacientes se ven beneficiados si se logra una detección precoz, ya que esto facilita también un tratamiento y diagnóstico tempranos, anticipándose así a una mayor progresión tumoral. En Estados Unidos, la mortalidad de cáncer de mama y colon ha disminuido en torno a un 40% en los últimos 25 años. Este descenso se ha logrado, entre otras cosas, como consecuencia de un aumento en los diagnósticos tempranos.

Tradicionalmente, el diagnóstico en cáncer se realiza a partir de muestras de tejido extraídas del tumor, un procedimiento conocido como biopsia. Las biopsias, sin embargo, necesitan ser realizadas y analizadas por personal especializado, lo que hace bastante complejo poder repetirlas, y casi imposible usarlas como estrategia de cribado. En este contexto, y como alternativa a las biopsias tradicionales, surgió el concepto de biopsia líquida, una técnica que consiste en el análisis de una muestra de sangre u otros

fluidos en lugar de la extracción directa del tejido afectado. Se pueden explorar muchos tipos de materiales biológicos diferentes a través de la sangre, y distintos estudios han asociado la mayor o menor abundancia de ciertas biomoléculas con la presencia de tumores, el estadio del cáncer o su mejor o peor pronóstico. Más concretamente, existe un gran interés en el caso de los miRNAs circulantes, porque son moléculas muy bien caracterizadas, abundantes en los fluidos corporales, y relativamente estables.

La secuenciación masiva es una estrategia común para proponer marcadores circulantes basados en miRNA a partir de sangre, suero, plasma o exosomas. En esta tesis, se han desarrollado y actualizado varias herramientas útiles para: 1) el procesamiento de experimentos de miRNA-seq para la obtención de perfiles de miRNA, patrones de expresión en los isomiRs y otras propiedades de los tRNA; y 2) realizar controles de calidad a experimentos de miRNA-seq mediante su comparación con más de 30.000 muestras obtenidas de repositorios públicos y procesadas uniformemente con sRNAbench. Finalmente, se desarrolló una base de datos curados a mano para albergar experimentos de fluidos con el objetivo de incrementar la reproducibilidad y reusabilidad de dichos datos.

sRNAbench es uno de los servidores web más utilizados para el análisis de experimentos de secuenciación de miRNA, con casi 50 mil trabajos lanzados desde su última publicación en 2019. Entre otros análisis, permite la obtención de perfiles de miRNA e isomiR, el mapeo a distintos genomas, y el uso simultáneo de varios índices. En 2019 se actualizó este servicio para incluir nuevos protocolos de preparación de librerías, especies y referencias de miRNA. Desde la publicación original de *sRNAbench*, la cantidad de datos de secuenciación de miRNA en GEO se ha triplicado. Por ello, se ha

adaptado el servidor para permitir la descarga automatizada desde *SRA* y el procesamiento simultáneo de múltiples muestras cuando provengan de un mismo protocolo. La herramienta de expresión diferencial, *sRNAde*, también ha sido actualizada para incluir más métodos y mejorar su interactividad.

Aunque el número de estudios de miRNA-seq en biopsia líquida ha aumentado considerablemente en los últimos años, los artículos y los datos disponibles públicamente no siempre se anotan en los repositorios de forma correcta ni uniforme. Como solución a esto, decidimos curar manualmente y procesar uniformemente 31 estudios disponibles en *SRA*, con un total de más de 1000 muestras de secuenciación de miRNA. Todo ello se organizó en una base de datos públicamente disponible, *liqDB*, que puede ser interrogada tanto para descargar cualquier conjunto de muestras como para generar o testear hipótesis. El conocimiento ganado a través de este proceso ha sido muy importante para desentrañar qué características determinan que una muestra de secuenciación de miRNA pueda ser considerada de buena calidad.

Frecuentemente, el control de calidad en los experimentos de miRNA-seq es un proceso que se pasa por alto. Aunque esta estrategia de secuenciación es muy común, los análisis aplicados para controlar la calidad no difieren mucho de los que normalmente se utilizan para un RNA-seq de RNA mensajero. Por ello, implementamos *mirnaQC*, la primera herramienta bioinformática especialmente diseñada para controlar la calidad de muestras de miRNA-seq. Para su desarrollo se aprovechó un algoritmo automatizado de detección y procesamiento de muestras, originalmente diseñado con la finalidad de poblar la base de datos *liqDB*, para analizar un total de

más de 36.000 experimentos de miRNA-seq. Basándonos en este corpus y en nuestra experiencia, propusimos un conjunto de 27 parámetros de calidad que se pueden explorar en su contexto mediante percentiles a través de los cuales los usuarios pueden realizar un control de calidad relativo de sus muestras. Además, el software permite seleccionar subconjuntos del corpus de referencia para que la comparación se efectúe con muestras similares a las de los usuarios. Uno de los conjuntos predefinidos lo forman experimentos obtenidos de fluidos corporales que de este modo pueden ser utilizadas como referencia para muestras de biopsia líquida. Hasta donde sabemos, *mirnaQC* es la primera herramienta bioinformática especialmente diseñada para el control de calidad de muestras de miRNA-seq y la única herramienta de control de calidad para datos de NGS que se basa en un conjunto de muestras para otorgar una evaluación comparativa.

Finalmente, para demostrar la utilidad de los métodos aquí presentados, los hemos aplicado a un conjunto de experimentos de miRNA-seq generados a partir de suero de pacientes con melanoma, un tipo de cáncer de piel que se origina en los melanocitos, las células que producen la pigmentación. Los potenciales pacientes de este tipo de cáncer se beneficiarían tremendamente de un diagnóstico temprano ya que las tasas de supervivencia varían mucho según el estadío de detección: la tasa de supervivencia a 5 años en los Estados Unidos es de un 99% en el caso de lesión localizada (Estadios 0, I y II), un 66% en el caso de metástasis ganglionar regional y solo un 27% en pacientes con metástasis a distancia.

Tras comparar el conjunto de muestras de controles sanos con el de los pacientes mediante varios métodos de expresión diferencial se encontraron una serie de miRNAs que aparecían de forma significativamente más

abundante en alguno de los dos grupos. También se encontraron miRNAs diferencialmente expresados entre los controles y los pacientes en estadios tempranos, lo que indica un gran potencial de los mismos como posibles biomarcadores de detección temprana. Finalmente, encontramos diferencias en los niveles de uridilización de algunos miRNA que correlacionan con la progresión de la enfermedad.

Palabras clave: miRNA, secuenciación, biopsia líquida, cáncer, melanoma, detección temprana, metástasis

Contents

Chapter 1

Introduction

1.1 Cancer

Despite the increasing amount of private and public funding entities investing in cancer research [1], there are many advances yet to be made in cancer diagnoses and treatment. As of 2020, there were 19.2 million new cancer diagnosis worldwide and 9.9 million deaths from all cancer types combined [2]. The initial section of this introduction will cover the definition and some basic aspects of the disease, the global relevance of cancer with a focus on Western countries and the importance of early detection and diagnosis.

1.1.1 The importance of cancer to society : a global health and economic issue

Trying to portray the importance of cancer to Western society may appear unnecessary since almost everyone reading this introduction has experienced, is experiencing or is going to experience some form of cancer during their lifetime, either as their own illness or as that of somebody close to them (35% of people now alive in Europe will be diagnosed with cancer by the age of 75 [3]).

Cancer is the leading cause of premature death (0-69 years) in Spain and in the vast majority of European Union (EU) countries and the second cause

for the remaining nations [2]. The same trend is observed in other Western countries such as the United States, Mexico, Canada, Australia, Brazil, Argentina, etc. and in further high or very high Human Development Index (HDI) countries such as Russia and Japan. Developing countries are less affected by cancer mainly because of a younger population and lower prevalence of other risk factors such as smoking, unhealthy diet and excessive body weight [4]. This is predicted to change by 2040, with rising cancer rates in low and middle HDI countries but still far from figures exhibited by high HDI nations today. In fact, 28.4 million new cases are expected in 2040 worldwide, a 47% rise from 2020 (19 million, 282,421 of which were in Spain), with higher increases in developing countries (64-95%) than developed ones (32-56%) [2] (Figure 1). These predictions are based on demographic transitions and risk factors that are expected to increase in growing economies.

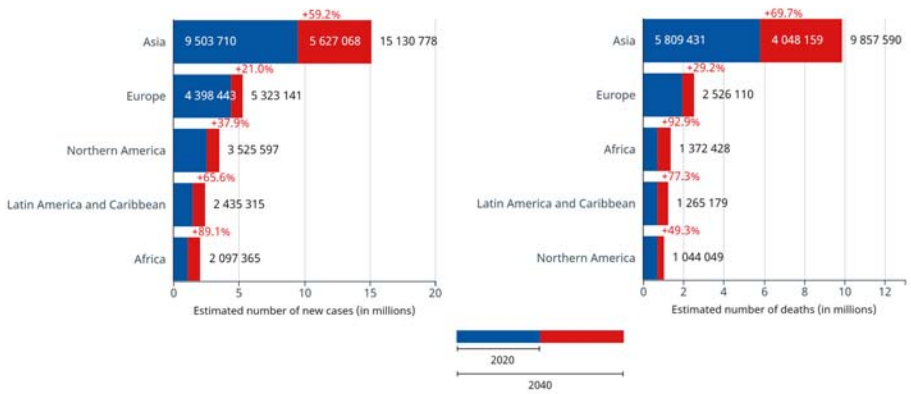


Figure 1: Estimated number of new cancer cases and deaths per year in 2020 versus 2040. Retrieved from https://gco.iarc.fr/, Global Cancer Observatory

Besides the dramatic consequences it can have in lives of people touched by the disease, there is also an economic burden on society as affected people will need healthcare, relatives may be provided with derived welfare or pensions and companies will have to temporarily or indefinitely replace ill or deceased employees. Furthermore, some patients will have to rely on family or friends for support during treatment or while they recover from the disease since they may find themselves unable to continue working. In a study from 2013 [5], Luengo-Fernandez et al. calculated that the economic cost of cancer for EU countries was on average equivalent to 102 euros per citizen every year, 4% of combined EU healthcare budget. They defined economic cost as total expenditure on healthcare and drugs, unpaid informal care provided by relatives or friends, lost earnings by premature deaths and costs of temporary or permanent cease of employment. Cost varied a lot depending on the country and the type of cancer analyzed but it correlated strongly with average income per citizen, which can be attributed to an increased cost of living and access to more expensive drugs and treatments. Similar countries in terms of GDP (Gross Domestic Product) per citizen could still display strong differences in average expenditure. Such was the case of Germany and the United Kingdom that had an almost two-fold increase in cost, more expensive in Germany, despite a very similar GDP. This fact highlights how public healthcare systems can be more affordable to deliver cancer treatment, although overall effectiveness of each system was not assessed in this work.

Cancer is also of great interest to the pharmaceutical and biotechnology industry. In 2019, 23% of all new drugs approved by the FDA were cancer therapeutics [6]. The latest estimations [7] put clinical oncology drug

development at the top of the list in terms of worldwide pharmaceutical investment in research and developement (R&D) with more than 92 billion of United States Dollars (USD) despite the current Coronavirus disease 2019 (COVID-19) pandemic. According to the same work, Oncology is predicted to remain on top by 2026, accounting for 22% of drug sales that year [7]. Companies justify this investment in two ways: first, most of the potential customers of their drugs live in wealthier countries so they can afford high prices for a unique potentially life-saving product and gain returns quickly; and second, new technologies and advancements initially made in the field of cancer research will eventually transition to applications in other diseases.

Such is the case of mRNA vaccines that have recently shown their success in helping to cope with (and hopefully eradicate) the coronavirus behind the ongoing pandemic. Several of the first approaches to achieve mRNA vaccines were attempted on cancer models [8] and the first relatively successful clinical trial of an mRNA vaccine displayed some effectiveness to generate antibodies against prostate-specific antigen (PSA) in metastatic prostate cancer patients [9]. In 2008 the first Phase I/II clinical trial of an mRNA-based vaccine proved to generate anti-tumor antibodies in melanoma patients, although it was not shown to be clinically effective [10]. Something similar happened with therapeutic monoclonal antibodies: even if cancer was beaten in the race to be the first target of this kind of therapy [11], many monoclonal antibodies to treat different types of cancer successfully obtained FDA approval since then [12, 13]. Mistakes and advances made during their development paved the way for a great range of diseases that can now be treated with more than 100 monoclonal

antibodies [14], though in the case of cancer they rarely are curative [15]. This gigantic research investment is coupled with great collaboration from the cancer patient community (around 5% of patients enroll in clinical trials [16] but 55% accept to participate when offered [17]) probably fueled in many cases by the lack of an appropriate treatment that can provide acceptable survival rates.

Interest in cancer research goes beyond the pharmaceutical industry. Public and private research funding entities also provide increasing resources to academic groups. In 2020, the National Institutes of Health (NIH) allocated ∼7 billion USD [18], 17% of their total budget, to cancer related projects and clinical trials (33% increase since 2013). Besides private, governmental, and academic funding, philanthropy has also arisen as a source of investment for cancer research. Good examples of this are several cancer projects, institutes and foundations started or financed by the Chan Zuckerberg Initiative [19] and the Bill & Melinda Gates Foundation [20] or Spanish counterparts like Amancio Ortega, who also donated diagnostic equipment to national hospitals among great controversy [21]. Despite allegations that these contributions were not altruistic but rather a campaign of reputation washing, they chose to fund cancer because of its relevance to society. In fact, 8 out of 10 American citizens surveyed in 2015 [22] supported medical cancer research and 74% were in favor of increasing federal funding. Voters were also 5 times as likely to support a president that prioritizes the fight against cancer, which doesn't come as a surprise since every age group surveyed put cancer as their top health concern and almost 9 out of 10 had met someone who had cancer (47% had a close friend or relative who currently has cancer). Richard Nixon

used this widespread view to boost his support by promoting the National Cancer Act of 1971 (although total NIH funding was decreased that year), an effort that has received bipartisan backing since then.

In summary, cancer is the leading cause of death worldwide, 10 million deaths in 2020, even amidst the current coronavirus outbreak (as of November 2021, 5 million reported deaths since the beginning of 2020), so there is ground to consider it an ongoing pandemic [23]. Furthermore, these figures are only expected to grow by 2040 [2] due to the world's aging population. This implies increasing challenges to healthcare systems around the globe, especially in developing countries. Western societies are aware of this issue and widely support current and even further public expenditure in cancer research [24], which is already one of the best funded topics. Finally, pharmaceutical companies make great investments in cancer research because of the still relatively low survival rates, the increasing number of patients worldwide (particularly in developed countries where new costly drugs can be afforded) and the anticipation that new technologies and therapies developed will be useful in drugs or products beyond cancer.

1.1.2 What is Cancer?

Cancer is defined by the WHO [25] as a large set of heterogeneous diseases that can arise by uncontrolled abnormal cell growth and multiplication of almost any cell type in our body. Cells with such a behavior are known as cancer cells and they tend to grow into a solid mass of tissue called tumor (except in the case of lymphomas and leukemias, cancers of the blood and bone marrow). So, contrary to the popular misconception, cancer is not a

single disease originating in different locations but rather more than a 100 different diseases that share some traits [26]. This is why adjusted survival rates 10 years after diagnosis can vary quite dramatically ranging from 1.1% for pancreatic cancer to 98.2% in testicular cancer [27]. Naming the cancer after the organ where it originated is common practice and can already confer a good overview of what kind of prognosis can be expected, although with some limitations. Perhaps one of the best examples of this limitation is small cell lung cancer and non-small-cell lung carcinoma, two types of cancers originating from the lung that have a 3.5 fold change in survival rate [28] (7-25%, in favor of the latter). Naturally, many other factors play an important role and some will also be covered in this introduction such as stage at diagnosis.

Potentially, tumor cells can also migrate to close and distant organs in a process known as metastasis, leading to major health problems beyond their area of origin. If cancer is certainly one of the most feared diagnoses by most patients, it is probably closely followed by metastasis. Metastasis is in fact the most challenging clinical complication of most types of tumors and the leading cause of death in cancer patients [29], at least when it comes to solid tumors (66.7%). Metastasis occurs at later stages of cancer and differences in survival rates described above can partially be explained by how frequently each type of cancer becomes metastatic. This, in return, is determined by several factors such as how fast this process happens in that particular case, how early it is/can be diagnosed and the range of treatments available for that type of tumor.

In their classic publication, The Hallmarks of Cancer [30], Hanahan and

Weinberg proposed six essential traits that are shared by all cancer cells and that are acquired through cancer development to induce malignant growth: self-sufficiency in growth signals, insensitivity to inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. Later on [31], they updated their list with two more traits: reprogramming of energy metabolism and evading immune destruction. Although different cancer cells may rely on these acquired capabilities with varying degrees, it is hypothesized that incipient cancer cells will progress towards a malignant state by developing each of these features in a multistep process. This would explain why cancer is more common in the elderly (about 60% of cancers occur in people age 65 or older [32]) as cells have had more time to accumulate these malignant features [30]. Additionally, this framework can also explain why some cancer patients seem to initially respond to a given treatment but end up showing drug resistance: the initial response is a result of the "loss" of one or several targeted hallmarks by the cancerous cell; however, since the remaining hallmarks are intact, it is only a matter of time for the tumor to end up finding an alternative pathway to recover all of its missing malignant features.

There is a great amount of scientific literature describing how tumor cells from different types of cancer can acquire these alterations and eventually become malignant. In the following section some brief background on this process is provided.

1.1.2.1 Tumor initiation: mutations and other causes of cancer

Every organ in the human body is meant for a specific function. Therefore, each organ is formed by millions of specialized cells that are programmed and differentiated for a specific function in that particular organ or system. Healthy cells are also programmed to undergo a planned death after a number of cell cycles in a process called apoptosis. This is a normal healthy process that ends the life of cells more than 50 billion times a day in an adult body [33] (which only accounts for $\sim$0.5% of the total cell count) and that can also be triggered by a series of factors including cell membrane damage, mitochondrial damage [34] and viral infection [35] or by other cells that recognize damaged cells as cancerous [36]. Evasion of planned cell death is one of the cancer hallmarks mentioned above and malignant cells can acquire this trait early on in their malignization process. Loss of this mechanism eventually leads to uncontrolled proliferation and allows the cell for accumulation of further hallmarks without dying. Evasion of apoptosis is typically acquired through mutation of tumor suppressor genes, a set of genes that regulate in different ways the appearance of defective cells and their replication. Although the inactivation of a tumor suppressor gene is not enough to onset a cancer cell, it is normally required that it loses its function either by mutation or other mechanisms. It should be noted that just this mutation won't be enough to turn a cell malignant so further mutations are required.

As mentioned above , cells typically acquire cancer hallmarks through mutation. Mutation is a modification of the genomic sequence that can be relatively small like replacing a nucleotide by a different one or affect larger

portions of the genome such as whole genes or chromosome fragments. Mutations happen randomly all the time as a result of DNA repair or imperfect pairing of bases during DNA replication since DNA polymerase enzymes add the wrong nucleotide an average of once every 100,000 times [37]. This may not seem like much but it would translate into more than 100,000 mutations per cell replication [38], and each of us has billions of cells. Luckily, these errors are fixed in 99% of the cases by a series of different mechanisms. Nevertheless after they escape this control and cell replication is finished, they effectively become mutations since the "new" cell has no way of telling which nucleotides are actually new (mutations) or belong to the original sequence. Therefore, normal mutation rates are healthy and no reason for concern as most mutations are neutral (i.e. not beneficial or detrimental for the individual). However, they can still result in the acquisition of cancer hallmarks and, eventually, in the arising of cancer. This is also why mutations tend to accumulate with age and consequently so does the rate of people suffering cancer [38]. If we consider this random process a raffle with a very low success probability following a binomial distribution, chances of winning the prize (meaning getting a mutation that evolves into developing cancer) increase with the number of times the individual plays (i.e. number of years lived). Very fortunate gamblers may never get diseased whereas unfortunate ones may get it as early as in their childhood without a specific given reason.

Although mutations happen all the time and are part of the random process already described, external agents can also increase the ratio of mutation. Here, some relevant examples are listed:

- **Tobacco smoking:** chemicals contained in tobacco such as benzene lead to an increased number of mutations in lung cells either by direct DNA damage or by preventing proper repair [39], even in involuntary smokers [40]. Furthermore, risk of other types of cancer like nasopharynx, stomach, liver or kidney is also increased as a result of smoking [40].

- **Exposure to chemicals:** many other chemicals have carcinogenic effects such as arsenic, asbestos or aromatics contained in gasoline [41]. This risk mostly affects occupationally exposed people or people exposed to low-quality construction materials.

- **Exposure to radiation:** high energy radiation such as ultraviolet light can directly damage DNA if the exposure is long and frequent enough. Melanoma and other skin cancers are commonly a result of exposure to sunlight [42] but virtually any cell's DNA can be damaged if sufficiently exposed to X-rays or gamma radiation.

- **Diet:** Red and processed meat consumption increase risk of colorectal cancer at least in part because they may contain N-nitroso compounds and sometimes aromatic substances generated at high temperatures [43], both of which are carcinogenic.

- **Alcohol:** many studies have highlighted the link between alcohol and digestive tract (cavity, pharynx, esophagus, stomach, colon, rectum) cancers [44]. Acetaldehyde, the first metabolite of ethanol (drinking alcohol), induces DNA damage and therefore mutations by its reparation [45].

Other factors can cause cancer without creating mutations. These factors can alter gene expression in a way that will provide some cancerous traits to the cell:

- **Viral infections:** different viruses can cause cancer including Epstein-Barr virus, Kasposi's sarcoma herpesvirus and HPV (human papillomavirus). The latter is an example of genetic gain of cancerous traits without mutation: this virus has two oncogenes and cervical cells infected with it will start to progress towards cancer [46].

- **Inflammation:** several studies have provided evidence of the essential role of inflammation in the tumor microenvironment for the progression of solid tumors [47]. Cytokines released by immune cells can trigger events of angiogenesis and proliferation among others [48]. Therefore, normal inflammation and inflammatory diseases also play a role in cancer initiation and progression.

- **Obesity:** different mechanisms like inappropriate insulin levels or hormones and inflammation are behind the well-established increased risk of more than 10 types of cancer among obese people [49]. People with excessive body weight also have worse prognosis in those same types of cancer, in part because of the difficulty of dosing chemotherapy [49].

- **Lack of exercise:** it has been epidemiologically established that low physical activity increases the risk of cancer independently of its link to obesity [50]. Risk to several cancer types is diminished [51, 52] by following appropriate exercise guidelines [53].

- **Hormones:** several hormones can promote cell proliferation and therefore induce cancerous behavior [54]. Particularly, gynecological cancers have been associated with abnormal levels of estrogens [55,56].

These external factors and mutagens rarely work in isolation, so the initiation and progression of tumors is usually a contribution of several of them. All of these events, however, work by altering gene expression in a complex and continuously evolving interaction with environmental factors. Consequently, cancer is a very complex disease and the same phenotype can be achieved through different mutations or genetic alterations that are generally obtained by one of the causes described above. Therefore, cancer can be generally conceived as a genetic disease where mutations are not inherited but rather acquired through life.

Nevertheless, different studies have shown an unneglectable inherited genetic component that accounts for 10-15% of all cancers [57]. Identical twins analyzed in this study [57] showed an increased risk of the same type of cancer compared to dizygotic twins (11-18% to 3-9%), which implies the existence of a major genetic component. Genome wide association studies (GWAS) have identified hundreds of single nucleotide polymorphisms (SNPs) linked to an increased risk of cancer; however, the amount of heritable risk is still very limited and, therefore, cancer genetic inheritability is largely unexplained [58]. An exception to this are hereditary cancer syndromes, a set of cancers that "run in the family" and are caused by one or very few mutations that, when inherited by an individual, can onset cancer. The variants that cause many of these syndromes have already been identified and they account for about 5% of all cancers [58]. Still, most variation

in cancer risk can be explained by factors that are not inherited [59].

1.1.2.2 Tumor progression

Through mutations, cells will start to proliferate beyond their normal rate growing into tumors. Before these neoplastic cells have acquired malignant properties (i.e. they cannot spread to close or distant organs) we still call them benign tumors or neoplasms. Pre-malignant lesions, such as dysplasia and hyperplasia, normally precede malignant invasive tumors [60]. Traditionally, it was generally accepted that cancer progression followed a linear pattern where cells would first acquire unlimited replication capabilities (tumor initiation). Then some of them would become malignant, meaning they grow faster and gain the ability to spread to other tissues. Finally some cells would migrate to other organs and start growing new tumors labeled secondary tumors (metastases). Therefore we use the term "progression" to define this stepwise fashion of malignization of tumors [61], even though some tumors may stay benign or never fully progress to a metastatic phenotype.

Plenty of evidence suggests that most tumors arise from a single cell which proliferates to form a neoplastic clone [62]. As neoplastic cells in the clone accumulate more mutations in key regulatory genes, a variety of clone sublines arise together with different phenotypes. This process is followed by a continuous selection of the most malignant cells which proliferate more and, therefore, outgrow other competing cells in a model termed "clonal evolution" [63]. According to this model, most new clones that continuously arise will be eliminated because of metabolic disadvantage or immunologic

recognition but occasionally one will have some kind of selective advantage and be established as the new dominant clone. As time passes, increasingly more aggressive subclones are selected, which normally leads to a more aggressive tumor.

Although tumors contain multiple subclones and consequently keep some of their heterogeneity, which can account for future drug resistance, with time a few clones that proliferate more rapidly take over most of the tumor. As a result, tumors become more malignant as they progress. This translates in cells displaying loss of differentiation, irregular shape and size plus a large nucleus and in the tumor increasing its growth rate and breaking the basement membrane [64]. The tumor will also increase angiogenesis to keep up with the growing population of cells and some will invade adjacent tissue.

Eventually, some cells will acquire a metastatic phenotype and migrate out of the primary tumor. Not all malignant cells that migrate will succeed in their effort to generate a secondary tumor but the ones that manage to establish a metastatic niche will grow forming a new clone. As a consequence, secondary tumors can have new properties compared to primary tumors since the metastatic environment can push clonal evolution in a different direction.

An schematic view of the described clonal evolution model is showed in Figure 2.

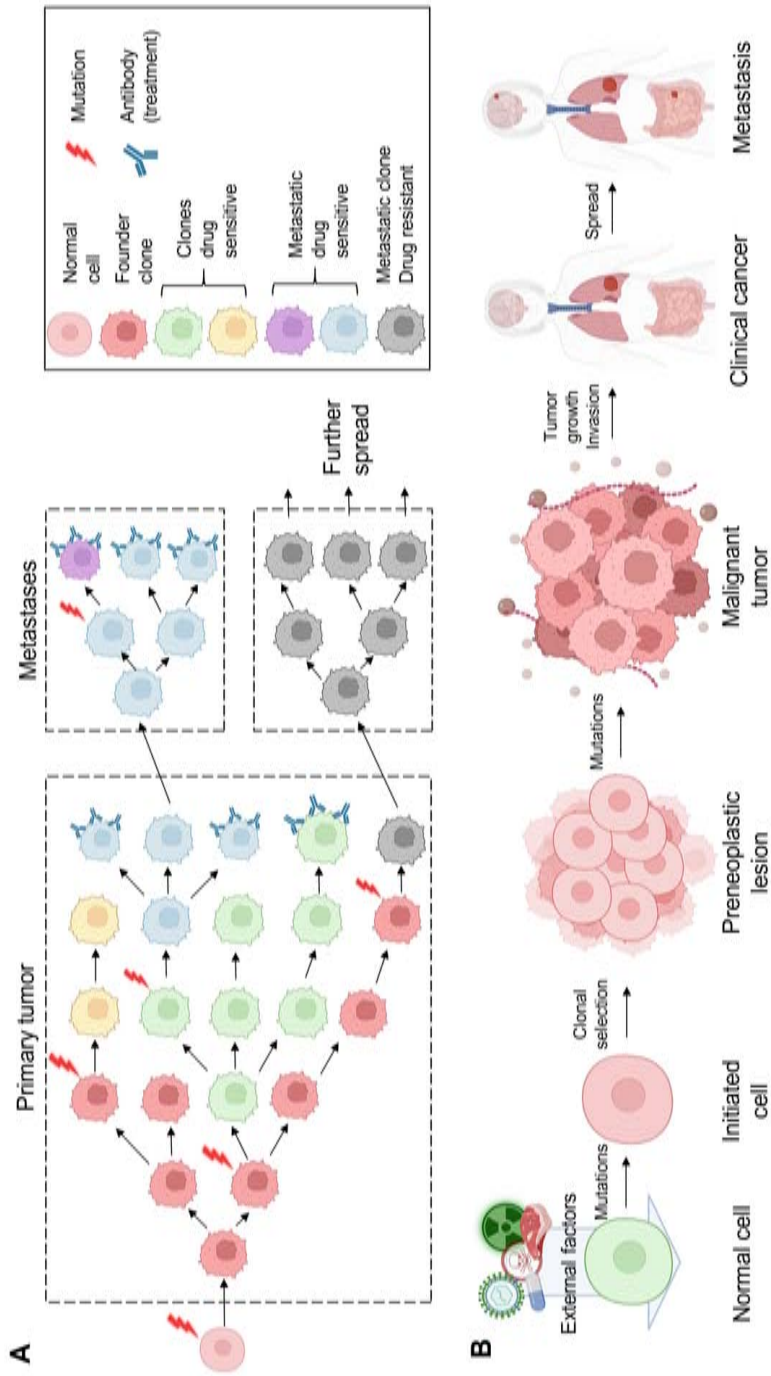


Figure 2: Cancer progression. (A) Schematic view of the clonal evolution model. Cells that survive the treatmen (grey) are positively selected *Adapted from [65]* **(B)**Schematic summary of tumor progression from a normal cell to metastatic spread

1.1.2.3 Cancer invasion and metastasis

A malignant tumor is defined by its capacity to invade close and/or distant organs. Metastasis is the final and most challenging outcome of cancer and responsible for most deaths in patients with solid tumors. Despite this, metastasis remains poorly understood [66]. A distinction can be made between cancer invasion, where tumor cells penetrate adjacent tissues, and metastasis where cells migrate to distant organs and establish secondary tumors. The former usually precedes the latter since the progression tends to proceed in an orderly fashion termed metastatic cascade [64]. This process is divided into five steps [64]: invasion of the basement membrane and cell migration; intravasation into the surrounding vasculature or lymphatic system; survival in the circulation; extravasation from vasculature to secondary tissue; and finally, colonization at secondary tumor sites.

Before they can migrate from their tissue of origin, cancerous cells must accumulate mutations that stop the expression of adhesion proteins such as E-cadherin and integrins that maintain healthy cells tethered in place [66, 67]. Once cells are detached from their adhesion to neighboring cells, they can migrate to other tissues. This happens in three distinct ways:

- **Invasion of adjacent areas:** Cancerous cells resort to built-in genetic programmes to invade and migrate surrounding areas. For instance, leukocytes also migrate as part of the inflammation response, a similar mechanism is used by cancer cells [67]. Groups of cells can also migrate together in a mechanism known as collective migration where a set of firmly interconnected tumor cells form local contacts

among them and degrade the extracellular matrix to create space for invasion. This type of migration recapitulates essential processes of embryonic development and wound healing [67].

- **Lymphatic system:** The most commonly occurring cancers spread first by lymphatic metastasis [66], where cancerous cells invade lymphatic nodes and ducts to migrate. This process occurs in stages: premetastatic invasion, approach, penetration, translocation, intranodal settling, growth and destruction of the lymph node and metastasis to further nodes [66]. As the tumor removal surgery happens it is common medical practice to biopsy or remove lymph nodes near the tumor in order to examine it for cancer spread. A node that shows signs of storing cancerous cells is then called a "positive node" and should be removed to prevent further progression. This form of localized spread is still not conceived as metastasis but the patient's prognosis worsens. The lymphatic ducts eventually return the lymphatic fluid to the bloodstream, releasing metastatic cells into the blood from where they can migrate further.

- **Hematogenous dissemination:** Metastatic cells can actively or passively reach [66] the bloodstream by directly intravasating into blood vessels around the tumor or *via* the lymphatic fluid in the fashion described above. Once a cell has reached the circulatory system, they need to survive long enough to eventually reach a distant site, which they do by interacting with cytokines in their microenvironment [66]. Cells can migrate alone or in clusters that also contain stromal or immune cells from the original tumor and tumor-educated

platelets. The latter arrangement has an increased advantage to survive circulation and establish a secondary tumor [68].

After reaching a new tissue through any of the routes described above, the cancerous cell still needs to adapt to new cellular surroundings (metabolism and angiogenesis) in the new tumor site and escape detection by the immune system [66]. They can do so by using hypoxia-inducible factors (HIFs), which adapt the genetic expression of genes related to invasion, metastasis and other cancer hallmarks [69].

Both ways of dissemination are not mutually exclusive. In fact, it is frequently the case that a correlation between lymphatic and hematogenous metastasis can be observed [70]. For instance, a study examining prostate cancer patients described that 84% of patients with positive nodes also showed some degree of hematogenous dissemination [71]. Similar observations have also been made in pancreatic cancer, ovarian cancer and head and neck cancer [70].

It is also worth mentioning that there is a propensity for each cancer type to seed in particular organs, as initially discussed by Paget after his thorough postmortem examination of breast cancer patients [72]. His initial suggestion has since been supported by a body of evidence, including experiments that demonstrate that metastatic cells show preferential adherence to their target tissue [72]. For instance, primary tumors from the kidney, breast, colon, bladder, head and neck, plus melanoma have a preference to metastasize to the lung [73]. The lung is the second most common target organ of metastasis which does not come as a surprise considering the huge vascularization of this tissue: 300 million capillaries with a cross-sectional

area of 400 cm2 which represents 50% of the whole human body [73].

For the sake of simplicity, tumor progression has been described here as a linear process where fully malignant cells raise out of a grown primary tumor. However, current evidence suggests that tumor and metastases progression can also occur in parallel, that is, independent progression of metastatic tumors arising from early disseminated tumor cells prior or at the same time as the growth of a primary tumor [74]. Many studies [75–78] concluded that given their growth rate and their size at time of detection, some metastases must be initiated before the primary tumor is diagnosed, as they were too large to have spread from a late stage of the primary tumor [76]. Additionally, 5-10% of all cancers detected in Europe and the United States are metastases of unknown primary tumors [79, 80] and 5% of breast cancer patients with small tumors (less than 2cm) also presented metastasis at the time of diagnosis [81]. These facts can only be explained if malignant cells migrated and established a metastases way before the primary tumor had a detectable size.

Finally, although not explicitly mentioned here before, the metastatic cascade can continue after a metastases has been founded, that is, multiple metastatic events from the primary tumor and additional metastasis can arise from previous metastases and grow independently [74]. According to the clonal evolution model, cells acquire metastatic properties before establishing a secondary tumor, so selection will maintain them at least as able to keep doing so [74].

1.1.3 Cancer staging

As it can be derived from this text so far, cancer is a very complex and heterogeneous disease, even within each type of cancer. This means that at time of diagnosis, cancer type is frequently not enough to accurately decide a patient's treatment or to assess their prognosis [82] and some information about the extent of the spread can be critical in this effort. Furthermore, a classification is also useful to stratify patients, which enables the comparison of their response in clinical trials. Different staging systems exist but it is standard practice to stage solid tumors using the TNM system [83], which uses an alphanumeric notation to describe the extent of the spread at diagnosis:

- **T (which stands for tumor):** this parameter describes the primary tumor, increasing numbers correspond with larger and more spread tumors. Therefore, T0 means no evidence of tumor, T1-4 refers to the size and extension of the tumor (specific for each cancer type) and Tx is used when the primary tumor cannot be assessed.
- **N (which stands for node):** this parameter describes the regional lymphatic node(s) close to the tumor, increasing numbers correspond with more positive nodes and further spread. Therefore, N0 means no metastasis in nodes, N1-3 refers to how distant the positive nodes are (where 1 is close regional nodes and 3 very distant nodes). Finally, Nx is used when the nodes cannot be assessed.
- **M (which stands for metastasis):** this parameter describes the presence of distant metastasis. M1 is used for metastasis to distant

organs and M0 for absence of metastasis.

For instance, a breast cancer codified as T1N1M0 means the tumor has a size between 0 and 2 cm, there are positive nodes close to the affected breast and no sign of metastasis.

Most solid cancers can also be classified into not-so-detailed stages that go from I to IV. This more general system can still give a good overview of cancer progression and it varies from one type of cancer to other:

- **Stage 0:** Also known as carcinoma in situ (CIS) at this stage neoplastic cells are present in the tumor but they are not malignant and they haven't invaded adjacent tissue. This is not considered cancer yet.
- **Stage I-III:** These stages already denote the existence of a malignant tumor (i.e. cancer). Higher numbers are used for larger tumor sizes and more spread, although the specific classification varies with cancer type. For some cancer types, each stage can be furtherly divided by adding letters after the stage number such as in IIIa, IIIb, and IIIc.
- **Stage IV:** At this stage there is distant metastasis so other organs are affected.

Further information can be included as part of the classification process. For instance, mutations on specific relevant genes can be included as part of the diagnosis (e.g. RAS positive stage III melanoma). The mutational profile has an impact on the response to treatment and a key determinant of first line targeted therapies [83].

1.1.4 The importance of early detection of cancer

As we have described so far, tumors tend to become more aggressive and complex as they progress. Among other things, this potentially implies a more challenging treatment because of the increased spread and heterogeneity, which may eventually lead to drug resistance. As a consequence, late diagnosis is associated with lower chance of survival and higher costs of care in most types of cancer where this factor has been studied [84–86]. In fact, stage-at-diagnosis is one of the best predictors of outcome for many tumors: one-year survival showed a major decrease when diagnoses happened at stage IV in breast, prostate and colorectal cancers and decreased at every stage for lung and ovarian cancer [86] (Figure 3). In the last 25 years, the US has experienced a $\sim$40% mortality decrease in breast and colorectal cancer in part as a consequence of achieving earlier diagnoses [84, 87].

Healthcare systems have therefore great interest in implementing different strategies to detect, diagnose and treat cancer at its earlier stages in order to increase survival rates and decrease the cost of treatment. Conservative estimates put the cost-savings at 17% of healthcare expenditure in the US for all cancers combined after implementing already-available effective screening solutions [88]. However, such approaches do not come without limitations. Given the probabilistic nature of tests and screening procedures, false positives and negatives are to be expected yet the positive predictive value needs to be relatively high for a successful test to be established. Furthermore, testing for cancer can prove quite challenging given the distribution of its probability of occurrence: chances of developing cancer over the course of a lifetime are relatively high but very low

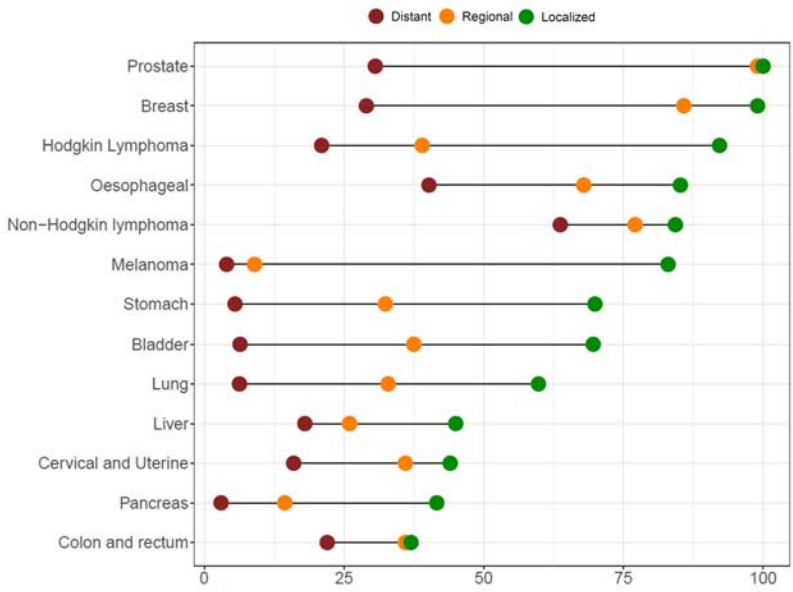


Figure 3: Five-year survival rate by stage at diagnosis for different types of cancers. Data from SEER 2011–2017, All Races, Both Sexes by SEER Summary Stage 2000.

at each particular moment of life [89]. For massive testing to make sense, these factors need to be taken into account. This is why most screening strategies are designed for high risk populations such as older individuals or those affected by comorbidities, where the prevalence is higher [89].

Despite the fact that these strategies have proved their efficacy in increasing survival rates and reducing treatment costs, most of the cancer research funding is allocated to late-stage treatments [88]. Nevertheless, new and developing technologies are increasing the precision of genetics-based screening tests which can revolutionize cancer diagnosis. Several companies such as Guardant Health [90] are integrating cheaper and more precise

detection technologies with a big data approach to improve detection of cancer in asymptomatic patients.

1.2 Liquid biopsy

Traditionally, extraction of tumor tissue is performed by surgeons in order to apply a series of tests to assess the extent of cancer progression, a process known as biopsy, a technique already described by medical encyclopedias from the medieval period in al-Andalus [91]. This technique allows histological definition and genetic characterization of the sample, information that can be very relevant to predict disease progression and response to therapy [92]. However, it does not come without limitations. For instance, the process is invasive and uncomfortable to the patient plus relatively expensive to the system (a highly trained specialist is required and the patient should first be prepared by medical staff) and it normally only allows for a single snap-shot sample of the tumor [92]. This inherent sampling bias can lead to underestimation of the tumor heterogeneity, among other shortcomings. Taking additional samples may not necessarily help as this is an inherently risky process for the patient who always faces a chance of having malignant cells metastasizing as a result of the procedure. Routine processing of biopsy specimens is also relatively slow so there is an inevitable delay between sampling time and start of the treatment.

These limitations motivated cancer researchers to look for new cheap easy non-invasive sampling techniques that could be applied multiple times without an increased risk for the patient.

1.2.1 The concept of liquid biopsy

Liquid biopsy is a medical testing approach that consists in sampling and exploration of bodily fluids as a surrogate for traditional biopsies. In short, the sampled fluid, most commonly blood, is tested for the presence of one or several of the following: circulating tumor cells (CTCs), circulating tumor DNA (ctDNA), exosomes, RNA or proteins [92] (Figure 4). Detection of circulating tumor DNA correlates with tumor burden and specific mutations can indicate the presence of drug resistant subpopulations that can proliferate despite therapy [93,94]. Combined with –omics approaches, this new testing paradigm has the potential to truly revolutionize cancer detection and diagnosis, in part due to its many advantages over traditional biopsies:

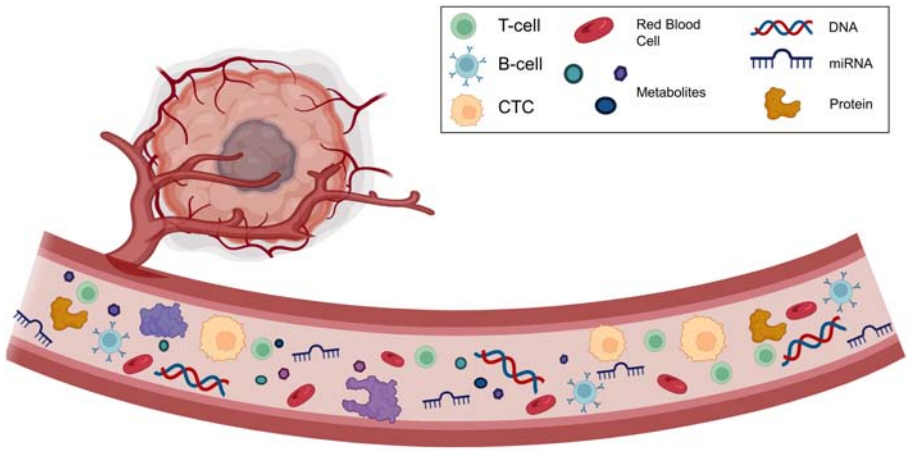


Figure 4: Circulating cells and substances that can be detected using liquid biopsy strategies.

- **Non-invasive:** a regular biopsy requires puncture or penetration of the tissue to sample it, which involves more risks, pain and recovery to the patient whereas sampling blood, urine or cerebrospinal fluid is almost painless and no recovery time is needed. This can also be a vital approach when the anatomical location is inaccessible or cannot be sampled for some other reason.

- **Early detection and diagnosis:** CTCs or genetic material can be detected before metastatic cells have established a secondary tumor. Furthermore, genetic changes happen before they translate into morphological changes, which means they can be detected before they could be diagnosed by an anatomical pathology specialist.

- **Inexpensive:** sampling blood or urine can be done by any trained nurse with very little equipment compared to regular biopsies that are normally performed by surgeons.

- **Easy:** little training, staff and equipment is needed for the sampling process, so it can be streamlined and incorporated in everyday practice more easily.

- **Fast:** sampling blood requires next to no preparation for the patient and it can be performed in just a few minutes. This means that the result is available faster and the test can happen shortly before the start of their treatment which may bring on crucial information considering how fast tumors can evolve.

- **Liquid biopsy enables multiple testing:** because blood sampling is inexpensive, easy and fast it can be repeatedly performed, which

could be highly impractical or sometimes impossible for a regular biopsy. This opens the door to different simultaneous applications like screening, monitoring or assessing the response to treatment. Resampling is also much more feasible if the initial sample is degraded or of insufficient quantity or quality.

- **Objective and precise:** morphological assessment of cancerous tissue is a key component of diagnosis in the traditional biopsy. Although necessary, these tests are harder to put in quantitative terms and have to be assessed by specialized staff, which can lead to unwanted human-caused bias and delays. Detection and quantification of genetic traits can be automatized and results have less room for interpretation.

- **Sensitive:** whether deliberate or not, complete sampling of the whole tumor is not always possible (it may even not be advisable). Consequently, important subclones may be missed in this procedure which leads to underestimation of the tumor heterogeneity. This can potentially be overcome by liquid biopsy approaches.

Despite the many advantages of liquid biopsies, tissue sampling is still more specific. Therefore, these two approaches would be most useful if used in combination rather than one replacing the other. Genetic information obtained on a first regular biopsy sample can be updated by several blood tests performed to assess response to treatment.

Although the concept of liquid biopsies was originally conceived to support cancer diagnosis and all examples brought up here will be related to

cancer, many other pathologies are being studied from this perspective including inflammatory and autoimmune diseases. The list of fluids sampled so far goes well beyond blood: urine, cerebrospinal fluid, amniotic fluid, fetal blood, breast milk, saliva, vaginal secretion, seminal fluid, bile, perspiration and menstrual secretion [95].

1.2.2 Biological materials analyzed by liquid biopsies

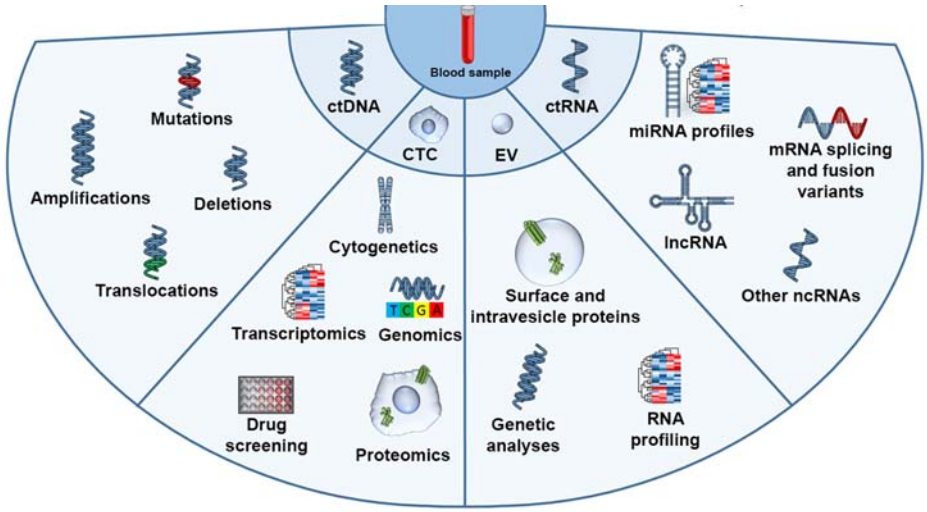


Figure 5: Circulating omics in human blood. Different analysis can be performed on cell free nucleic acids (ctDNA, ctRNA), Circulating Tumor Cells (CTC) or extracelullar vesicles (EV). *Adapted from [96].*

1.2.2.1 Circulating tumor cells

As described in a previous section, tumor cells can actively or passively shed from the primary or metastatic tumor and end up in the bloodstream as part of the tumor progression process, receiving the name of circulating

tumor cells (CTCs). CTCs can then "seed" at distant organs and initiate metastases [97]. Compared to the rest of blood cells, CTCs are very rare but their cancerous morphologic properties can be used to detect and isolate them from the bloodstream. Once cells are isolated using functional, immunological or physical techniques [98], analysis can get as complicated as required: from simply counting the cells, since the number of CTCs has been described to increase with tumor burden [99], to single-cell sequencing as a way to get a very detailed snap-shot of the mutational landscape or genetic expression. Furthermore, some studies [100, 101] have shown that CTCs reflect tumor heterogeneity, so characterizing them could be of great help to stay ahead of drug resistance. Along the same lines, a decrease in CTCs is associated with better response to chemotherapy in breast [102], lung [103] and prostate [104] cancer. Number of CTCs also correlates with progression-free survival and overall survival, so it seems a good predictor of outcome and indicator of relapse [105, 106].

Although not as developed as the analysis of CTCs, where the FDA approved [107] a CTC-based test to detect circulating cells in order to assess prognosis of breast, prostate and colorectal cancer patients; there are also great promises in developing tumor organoids from CTCs [97]. Tumor organoids are tiny models of tumors grown in a 3D semisolid matrix that resemble some of the tumor properties and can be used to predict tumor evolution and to perform drug screening [108]. CTC-derived organoids have already been used to characterize molecular properties of metastases in prostate cancer [109]. The resulting model maintained genomic alterations very similar to the parent tumor even after months of culture. Models derived from CTCs are more practical than regular tumor-derived organoids

since they do not require a biopsy of the metastatic tissue.

1.2.2.2 Extracellular vesicles

Extracellular vesicles (EVs) are naturally occurring bilayer-delimited particles of nanometer scale [110] that are released by cells and that can be found in several body fluids, which makes them perfect candidate targets for liquid biopsy tests. According to their size and cellular origin, EVs can be classified as exosomes, microvesicles or apoptotic bodies [110]. Because of the cellular origin of exosomes (lysosome), some initially attributed to EVs a waste disposal function. Nevertheless, their role in cell signaling and antigen presentation has extensively been studied [110]. EVs can have different cargo or contents, and their composition depends on the cell of origin although it is not necessarily similar to the contents of the cell that they come from [110]. Possible cargo includes proteins, nucleic acids such as DNA or RNA, lipids and metabolites.

Exosomes have been reported to play a role in virus transmission, which is relevant for certain types of cancer [111]. Furthermore, exosomes can also promote tumorigenesis, metastasis [111], immunosuppression and angiogenesis and they do so by transferring bioactive materials that have the intended effect in the recipient cell. Therefore, cancer cells can use this mechanism to escape immune surveillance and induce immune tolerance by manipulating the tumor microenvironment [112]. On the other hand, immune cells can also use exosomes to inhibit tumor growth and metastasis. The fact that they are used by both cancer cells and the immune system makes it a very interesting target of analysis since their content can reveal

mechanisms put in place by both sides of the fight.

Many studies have employed or are currently developing exosome detection strategies in liquid biopsy tests in saliva [113], urine [114], cerebrospinal fluid and blood [115,116]. Clinical application of this liquid biopsy strategy is only starting but some companies are already commercializing tests based on circulating exosomes such as ExoDx Lung, which uses a specific protocol based on qPCR to interrogate EGFR mutations in exosomal RNA [117]. Additionally, many clinical trials are currently evaluating different liquid biopsy tests based on exosomes [112].

Exosomal miRNAs deserve especial mentioning in this section as they are a particularly studied set of non-coding genes that are frequently proposed as biomarkers [118–120], more so because the methods and data developed, used and presented in this thesis are mostly centered around miRNA expression (although not necessarily associated to exosomes).

1.2.2.3 Circulating tumor DNA

Tumor cells release DNA into the bloodstream as they die from necrosis or apoptosis. Some have also proposed that this is a metastases development mechanism that works by transfecting neighboring cells [121]. Circulating cell-free DNA (cfDNA) also comes from healthy cells, the amount of tumor specific DNA (ctDNA) having been described to range from 0.01% to more than 90% [122–125], and it can be identified by detecting tumor specific mutations typically by whole genome sequencing (WGS). Nevertheless, cfDNA is frequently used as a surrogate for ctDNA and higher plasma levels have been found in patients from different cancer types [126–128].

Interestingly, ctDNA is detectable at early cancer stages but the size of the tumor does not correlate with the detected DNA concentration [129]. As a consequence, ctDNA can be used for early detection of cancer but progression should be assessed by other analysis including mutational profiling. Still, there is a chance that DNA amounts can be used to confirm disease-free status if ctDNA levels drop significantly after surgery [130].

A series of clinical studies have described successful applications in cancer detection, prognosis, recurrence prediction or progression assessment by means of targeted or untargeted tests [131–135]. Targeted approaches aim to interrogate specific genes that are known to mutate with cancer or chromosomal regions that are frequently translocated in tumor cells [136]. Successful applications of this approach include a method to detect mutated KRAS, a proto-oncogene, by digital PCR (dPCR) in colorectal cancer patients [135]. More sophisticated broader techniques allow for the sequencing of targeted regions [133], which is advantageous compared to single gene assessments since most tumors will actually not carry the most common mutation or may only acquire it at a later stage [133].

Most widespread untargeted approaches include some form of the abovementioned WGS which can assess all possible mutations presented in the sample without any prior knowledge at a competitive decreasing price. This powerful technique comes at a cost though, as it requires proper bioinformatics analysis to successfully process the sequencing files and call true mutations. Once a mutational profile has been achieved, point mutations must be tracked down in available databases or resources to decipher their importance. It should be noted that the overwhelming majority of truly

called mutations or SNPs will be meaningless or not related to cancer itself.

1.2.2.4 Circulating tumor RNA

Contrary to what happens with DNA, RNA is quite unstable and susceptible to RNase-catalyzed degradation [137], which would on paper discard them as valid biomarker candidates. Nevertheless, circulating cell-free RNAs (cfRNAs) seem to be protected inside microvesicles or ribonucleoprotein complexes in order to avoid degradation [138]. In fact, several protein-coding and non-coding RNA (ncRNA) molecules have been confirmed to be detectable in blood and potentially appropriate liquid biopsy biomarkers for different types of cancer [137, 139–142].

Although virtually all RNA subtypes have been found in circulating blood, including mitochondrial RNA, bacterial RNA and other foreign RNA [143], most RNA-based liquid biopsy studies have overwhelmingly focused on analyzing circulating small ncRNAs, particularly miRNAs. This is probably due to their well-characterized relatively high abundance and stability in most bodily fluids [144]. Still, several studies have confirmed the utility of detecting mRNA biomarkers for liquid biopsy tests [137, 145] and specific sequencing protocols allow for detection of mRNA fragments [146].

Different types of non-coding RNAs such as long non-coding RNAs (lncRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA), circular RNAs (circRNA) and yRNA have also been detected in several bodily fluids. Since most studies focus on miRNAs, the rest of RNA species are normally co-detected as products of the small-RNA sequencing protocol, an adaptation of RNA-seq that targets RNAs in the

range of mature miRNA nucleotide length [147] (this protocol is covered in the next section of the introduction). MiRNA is often the preferred source of circulating biomarkers because of their relative abundance and stability in biofluids since they are resistant to endogenous RNases, low or high pH, extended storage periods, boiling and several freeze-thaw cycles [148] and it has long been established that changes in miRNA expression correlate with progression in multiple types of cancers [149–151]. This outstanding resistance can partially be explained by their binding to the Argonaute family of proteins and to high density lipoprotein complexes [152, 153].

Since circulating miRNA are gaining a lot of attention as potential biomarkers, it is not surprising that many studies have already proposed or even established blood-/serum-/plasma-based liquid biopsy tumor markers for the assessment of diagnosis, prognosis or response to treatment in both solid and hematological tumors [154]. Several panels discovered using different high-throughput technologies have been developed for breast cancer [155, 156], colorectal cancer [157, 158], glioblastoma [159], hepatocellular carcinoma [160], lung cancer [161, 162], pancreatic cancer [163, 164], ovarian cancer [165, 166] and melanoma [167–171] along with others [154]. Most shockingly perhaps, is the case of melanoma where no proper classical tumor marker is available and disease can only be staged after a traditional biopsy [172]. Researchers in a handful of studies [167–171] were able to propose panels based on a small number of miRNAs to separate healthy subjects from melanoma patients and stratify them according to cancer progression.

1.2.2.5 Other biological materials

Materials described until now have long benefited from untargeted high-throughput detection technologies, which made them suitable choices for the biomarker-discovery phase of studies. Nevertheless, evolving technologies have made it possible to include more –omics to the high-throughput trend. Next, the use of proteomics and metabolomics in liquid biopsy tests is briefly discussed.

Proteins are responsible for most cell functions and their expression deregulates with cancer. As such, they can be very useful biomarkers. In fact, most classic cancer biomarkers are proteins including estrogen receptor (ER), prostate-specific antigen (PSA), alpha-fetoprotein (AFP), human chorionic gonadotropin (HCG), CA 19-9 (cancer antigen 19-9), CA-125 (cancer antigen 125), CD30 and CD20. These and other proteins can be measured in blood to help in diagnosis, monitor response to therapy or in tissue samples to determine molecular subtypes. Traditionally, each protein had to be individually assessed by measuring their activity or be targeted by monoclonal antibodies. Until recently, this has delayed high-throughput proteomics liquid biopsy studies as systematically measuring each target was impractical. Nevertheless, studies based on panels of proteins have been developed to assess different aspects of the disease from blood, urine and saliva, sometimes in combination with other molecules [173–175]. Recent technological developments in the field of proteomics, like mass spectrometry (MS)–based high-throughput proteomics [176] or direct protein sequencing by nanopore [177], are making it possible to advance towards higher throughput proteomics studies. Examples of successful applications

of this approach include the use of circulating VEGF to predict response to treatment and overall survival in advanced melanoma patients [178] and a panel of biomarker proteins to detect bladder cancer from blood samples [179].

Regular blood and urine tests typically report single metabolites in their panels including glucose, urea, uric acid and creatinine among others. Metabolomics however, attempts to account for the whole set of metabolites present in a sample, which can be done in a fast cost-effective way by means of a nuclear magnetic resonance (NMR) spectrometer or mass spectrometer (MS) [180]. The cancerous cell has to adapt its metabolism to continue rapidly growing and replicating even in hypoxic or insufficient nutrient conditions. Consequently, a great number of metabolites have been reported to significantly change between tumors and corresponding healthy tissue [180]. There are also numerous examples that successfully implemented this approach in several biofluids such as blood [181] and urine [182].

1.3 miRNAs (and other small non-coding RNAs)

miRNAs are endogenous small ncRNA molecules of ∼22 nucleotides that mediate translational repression of target mRNAs through antisense base pairing [183]. miRNAs have been detected in all bilaterian animal species and many of them are evolutionarily conserved [184]. Their deep and widespread presence in animal genomes hints an important regulatory influence in a wide variety of physiological processes. In fact, mice studies have shown that disrupting miRNA genes often leads to defects in the develop-

ment of individuals [183] and mutations or aberrant expression are associated with cancer, immune, cardiovascular and neurological disorders [185]. Additionally, as it has been described above, miRNAs are secreted into extracellular fluids, potentially making them ideal biomarkers.

In the previous section the relevance of miRNA-based liquid biopsy approaches has already been highlighted. In the current section I will present some facts about miRNA, their function and their biogenesis, together with a next generation sequencing (NGS) approach, termed miRNA-seq, to detect and quantify them. Because this sequencing protocol relies on size-selection of inserts in the range of mature miRNAs, other small RNAs, which I will also describe here, are normally co-detected. Finally, the computational processing and quantification of miRNAs and their isoforms (isomiRs) from sequencing data will be addressed.

1.3.1 miRNA discovery and function

MiRNAs were initially discovered by molecular geneticists in C. elegans before the first draft of the human genome [186]. These researchers described that a non-coding gene, lin-4, produced a short 22 nt RNA that had partial imperfect complementarity to 3'UTRs (3' Untranslated Regions) creating a temporal decrease in the target protein. A similar gene also discovered in *C. elegans*, let-7, was later described in human and further animal species [187] where it conserved its function and expression pattern. Since then, sequencing technologies have fueled the discovery of many miRNA genes. The latest miRBase release [188], a database that collects miRNA sequences and attempts to provide them with consistent

naming, contains 48 860 mature miRNA annotations of which 2654 belong to humans. There is good reason to believe that these numbers are an overestimation, especially in the case of human miRNAs [189]. Nevertheless, conservative analysis still puts the number of human miRNA genes in well over 500 [183] and preferentially conserved interactions with most mRNAs have been described [190].

Most frequently, miRNA binding to specific complementary sequences present in the 3'UTR of target mRNAs downregulates gene expression by either promoting mRNA degradation or preventing the mRNA from being translated. However, interaction with other genomic regions including promoter, 5'UTR or coding sequence (CDS) has also been described [191]. Additionally, there is increasing evidence that miRNAs can also posttranscriptionally stimulate gene expression by both direct and indirect mechanisms [192].

A 6-8nt sequence that starts at the 2nd base from the 5'-end of the mature miRNA transcript is the most important determinant of target recognition [193]. This region is known as seed and base-pairing between the miRNA and the mRNA needs to happen for downregulation of the target [193]. miRNA genes that share a seed sequence are therefore expected to have overlapping targets and consequently have similar effects in cell physiology, which is why miRNA genes are grouped into families using their seed region. Nevertheless, this information is not enough to reliably predict all functional targets of a miRNA [194] and accurate miRNA target prediction based on sequence composition still remains somewhat unsolved.

1.3.2 miRNA biogenesis

MiRNAs first precursors, known as pri-miRNAs are transcribed by RNA polymerase II, the same enzyme that transcribes mRNA transcripts [195]. The pri-miRNA then forms a hairpin with a loop domain, a structure that is recognized by Drosha, an endonuclease in the nucleus of the cell [196]. Drosha cuts each stem of the hairpin with a 2bp offset to produce the pre-miRNA, a ∼60nt hairpin. This hairpin is then exported to the cytoplasm by Exportin 5 and subsequently processed by Dicer, another endonuclease that cuts both strands of the pre-miRNA to generate the miRNA duplex [197]. The resulting miRNA duplex is a double stranded RNA formed by the mature miRNA (or guide strand) and the passenger strand (miRNA$*$) with a characteristic ∼2 nt 3' overhang on each strand end as a result of Drosha and Dicer processing [197]. Finally, the duplex is loaded into a protein from the Argonaute family and the passenger strand is ejected to allow the recruitment of the rest of the RNA-induced silencing complex (RISC) [198]. Once formed, this complex is responsible for recognition of and binding to targets [198]. Cleaving of the target RNA can happen if there is perfect base-pairing but is rare [183].The miRNA biogenesis process is summarized in Figure 6.

1.3.3 miRNAs role in cancer

Since miRNAs are involved in virtually all physiological processes, it is not surprising that their dysregulation is associated with the development of several pathologies including cancer. As a matter of fact, loss of or improper expression of miRNAs has been described by many studies as a cause or

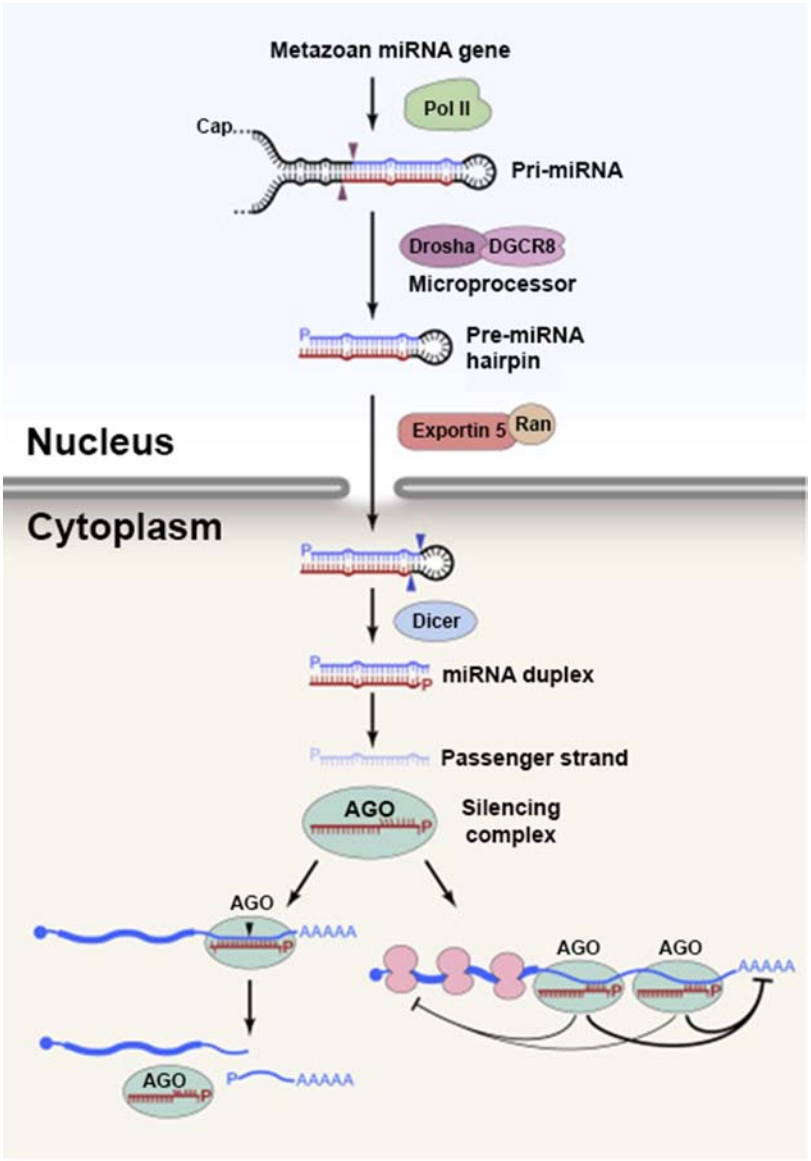


Figure 6: Biogenesis and action of microRNAs. *Adapted from [199].*

consequence of carcinogenesis [200]. An abnormal miRNA expression has been shown to impact cancer hallmarks leading to increased growth rates, proliferation, activation of invasion and metastasis and angiogenesis [200]. In that sense, some miRNAs behave like tumor suppressor genes or oncogenes.

Already in 2002, researchers described that miR-15a and miR-16-1 were located in the 13q14 chromosome region, a DNA fragment frequently deleted in B-cell chronic lymphocytic leukemia (B-CLL) cells [201]. Both miRNAs were found to be downregulated in most B-CLL cases. Later on, miR-16-1 was described to work as a tumor suppressor by repressing the expression of Bcl-2, an anti-apoptotic protein overexpressed in some malignant solid tumors [202]. In mice, experimental deletion of these two miRNAs provoked B-CLL-like phenotype, which demonstrated the role of these in tumor suppression [203]. Deletion of tumor suppressor acting miRNAs is one of the ways cells can acquire cancerous properties.

Similarly, miRNAs can also behave as (proto-)oncogenes, i.e. normal genes involved in the promotion or regulation of proliferation or inhibition of apoptosis and that can increase their activity by means of different genetic alterations [204]. The miR-17–92 cluster was observed to be amplified or translocated in B-cell lymphoma [205], lung cancer [206] and T-cell acute lymphoblastic leukemia [207]. These genomic alterations led to increased expression of the miRNA cluster and increased proliferation. GWAS later revealed that many miRNAs are located in genomic hotspots associated with cancer. This suggests that aberrant miRNA expression can arise from deletion or amplification of specific chromosomal regions that

contain miRNA genes, leading to carcinogenesis [200].

Many studies have since highlighted some role of specific miRNAs in the development of different types of cancer and proposed biomarkers for detection, diagnosis or prognosis as well as miRNAs as therapeutic targets or tools [200].

1.3.4 miRNA sequencing (miRNA-seq)

Before the arrival of NGS technologies, identification and characterization of miRNAs heavily relied on traditional molecular techniques to clone and sequence individual small RNAs [208]. Compared to current approaches, individual sequencing was slow, labor intensive, low-throughput and expensive. In 2007, Lu et al. [208] described their adaptation of an RNA-seq (RNA sequencing) protocol to sequence small RNAs. This protocol was initially developed with Arabidopsis samples but it worked just as well with animal RNA. Together with improving bioinformatics analysis, this method facilitated a boom in miRNA discovery and research. Today, more than 36,000 miRNA-seq (also termed small RNA-seq since other small RNA molecules are also generally sequenced) experiments are publicly available on SRA (Sequence Read Archive) [209], the largest public repository of sequencing data.

Although different variations and commercial kits of this protocol are available, in the next paragraphs an overview of the basic wet-lab and bioinformatics methodology to analyze miRNA through sequencing is provided.

1.3.4.1 Library preparation and sequencing

Although many small RNA-seq protocols based on different commercial reagents are available, most rely on similar or equivalent processes that result in cDNA libraries that are suitable to be sequenced by Illumina's machines. Here, I will describe a general protocol heavily based on Illumina's "TruSeq Small RNA Library Prep Kit".

1. **Total RNA isolation**

   Total RNA is the input material needed for this protocol but it is not included as a step on it. Initial samples can be either tissue or fluids which need different chemical treatments to isolate their RNA. Tissue samples should be snap-frozen, grinded to powder and then resuspended before application of TRIzol (a monophasic solution of phenol and guanidinium isothiocyanate, the most widely used RNA extraction chemical) [210]. For fluids, TRIzol can be directly applied on the sampled volume. After mixing and solubilization of the chemical and centrifugation, two phases will form. The upper phase will contain the RNA and then after several steps of precipitation and re-extraction the RNA is stored in purified ethanol at -20^{o}C [210].

   Several commercial kits claimed to outperform TRIzol are available for targeted extraction of miRNAs, particularly from fluids. These methods rely on additional chemical properties of miRNAs (for instance, miRNeasy uses a silica membrane column so that longer RNAs get "trapped" and are removed from the solution). A similar approach is employed by mirVana Isolation Kit that uses a glass fiber

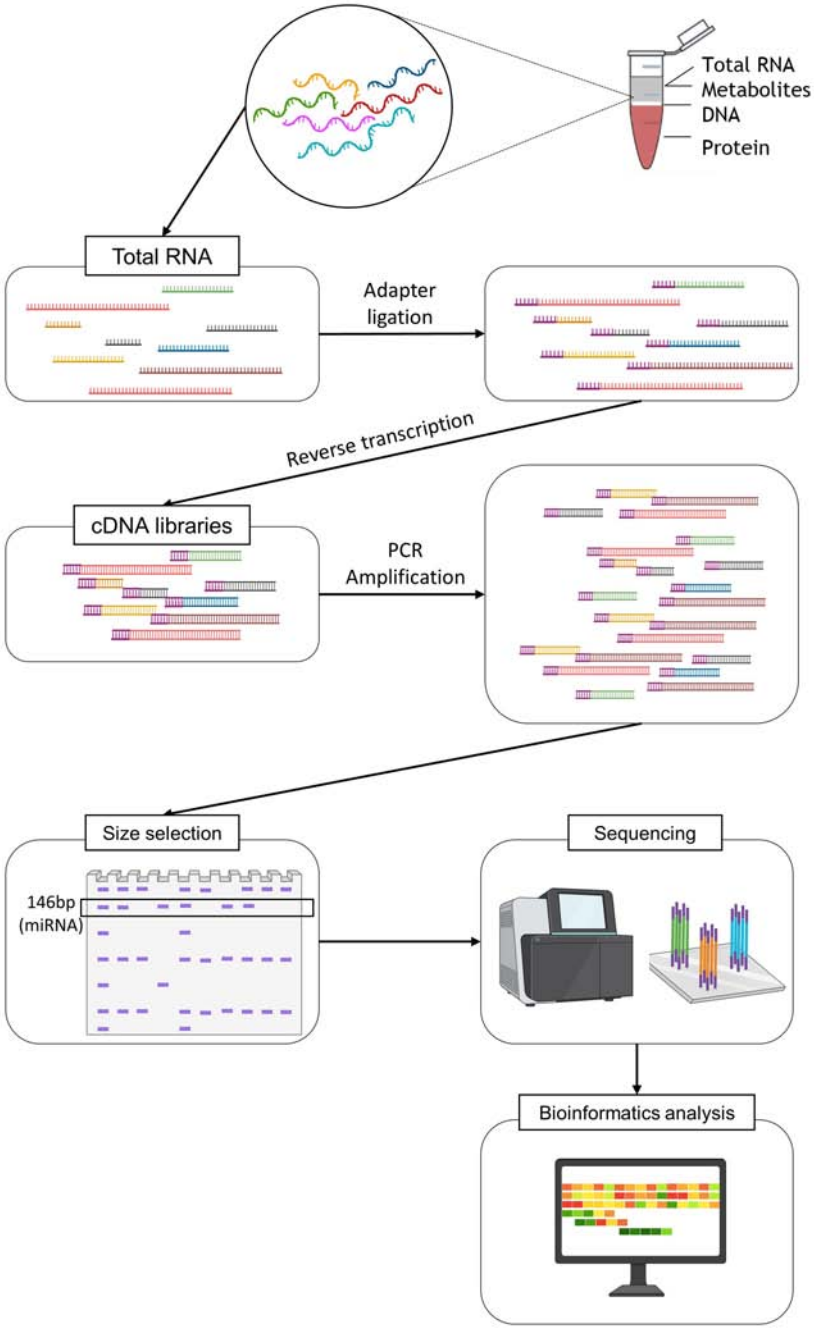


Figure 7: Small RNA library preparation and sequencing from isolated total RNA.

filter to achieve similar effects. Modifications of these kits are also commercially available to obtain miRNAs from exosomes and other EVs. Several studies have compared the impact of using each protocol on plasma samples [211–213]

2. **Adapter ligation**

   Although more recent technologies can directly sequence RNA, NGS works from DNA. Before a retrotranscription step to convert RNA to DNA can be introduced, 3' and 5' adapters have to be ligated. These will act as binding sites for reverse transcription and for PCR primers for amplification. The adapters and protocol are designed to take advantage of the 5' phosphate group characteristic of miRNAs to avoid sequencing degradation products. Normally, a T4 RNA Ligase 2 mutant, which shows some preference for given sequences, is used to ligate the adapters. To avoid this potential source of bias, overnight ligation (instead of 2h) has been proposed. Provided enough time, the enzyme will exhaust all RNA available even if some fragments tend to be ligated first.

3. **Reverse transcription and PCR amplification**

   A reverse transcription step generates DNA molecules from adapter-ligated RNAs. This is followed by PCR amplification, a process to increase the amount of DNA that selectively enriches fragments with adapters at both ends. This is achieved by annealing of two primers specifically designed to bind sequences present in the adapters. Index tags are introduced in this step. This allows for pool sequencing of several samples, which makes the process higher-throughput and

consequently cheaper.

PCR amplification however, seems to be a great source of bias in sequencing-based miRNA quantification [214]. Different approaches have been designed to reduce this unwanted effect like the use of random adapters [215] or the introduction of unique molecular identifiers (UMI), index-like tags that can be used to collapse reads generated from the same RNA fragment avoiding overestimation.

4. **Size selection**

   Because we are only interested in sequencing a specific range of RNA lengths, a size selection step is needed. Originally, this step was performed before reverse transcription and amplification to enrich for desired RNAs but more recent protocols have moved it back to when DNA is available since it is more stable to work with.

   In this step, DNA libraries are run in a gel electrophoresis to separate them according to their lengths. Once separation has been achieved, bands corresponding to 22 and 30nt (147 and 155nt after amplification) should be cut out with a razor blade and kept. Which small RNAs are sequenced and their coverage varies depending on the band selection. The purified libraries can be recovered from the cut-out bands by chemical treatment and resuspension in ultrapure water.

5. **Library check**

   Libraries can be run on a Bioanalyzer (Agilent Technologies) using a DNA chip to check the amount and purity of RNA at different

lengths. A distinct peak should be observed around 147-150nt, which corresponds to the length of mature miRNAs. This step is performed as a quality check of the libraries.

6. **Library normalization**

   Libraries concentrations are normalized so they can be pooled together for sequencing. Samples can be stored after this step until sequencing happens.

7. **Sequencing**

   Several sequencing technologies are available but the most widespread is Illumina's sequencing by synthesis, which I will briefly describe. Inside the sequencing machine, DNA fragments are captured by the surface of the instrument, which is covered in oligos that are complementary to the adapter sequence and subsequently amplified in a process termed cluster generation. After clusters have been generated, the sequencing by synthesis step comes next. Using modified nucleotides that include laser-detectable fluorescent tags, a complementary strand is generated from the cluster template. The process is designed so that only one nucleotide can be added at a time, allowing the laser in the instrument to measure the fluorescent dye at the end of the cycle [216]. These measurements are recorded for each cluster and stored by the machine as reads. This whole process results in a data file (fastq) that contains the sequences of the reads, including adapters and possibly unwanted artifacts plus quality parameters derived from the fluorescence detected for each nucleotide in the read.

1.3.4.2 Bioinformatics data analysis

After the sequencing process is completed, we are left with a fastq file containing reads and a Phred score per nucleotide, a measure of the quality that describes the probability of a base-calling error [217]. However, no indication is available of miRNA sequences or abundance. In order to achieve this, bioinformatics analysis have to be applied. In summary, reads have to be preprocessed, aligned to a reference and subsequently quantified.

Several miRNA-seq analysis pipelines and tools are available but they mostly rely on similar or equivalent steps with slight variations. Next, an overview of the quantification processed, mainly based on *sRNAbench* [218] is described.

1. **Read pre-processing: trimming and collapsing**

   As mentioned above, RNA fragments are ligated to adapters in order to allow sequencing, therefore these need to be removed from the read for proper identification of the fragment sequence. Additionally, each position of every read is checked for low quality nucleotides. Low quality read endings are trimmed together with adapter sequences and short reads are discarded.

   Unlike regular RNA-seq, mature miRNAs tend to be sequenced in their integrity or close. This means that a read is likely to capture the whole miRNA sequence or at least most of it. Furthermore, considering the relatively short length of mature miRNAs and that shorter reads are discarded, there is not a wide range of possible sequences derived from the same miRNA that can be detected. Additionally,

a few sequences tend to capture the majority of the sequenced reads assigned to a miRNA, and a few miRNA genes will make up most of the miRNA in a cell , sometimes up to 90% [219]. Typically, miRNA tools take advantage of this fact by collapsing all identical reads to dramatically reduce the number of alignments (Figure 8). The count of reads must be kept though, so they can be considered in the quantification step.

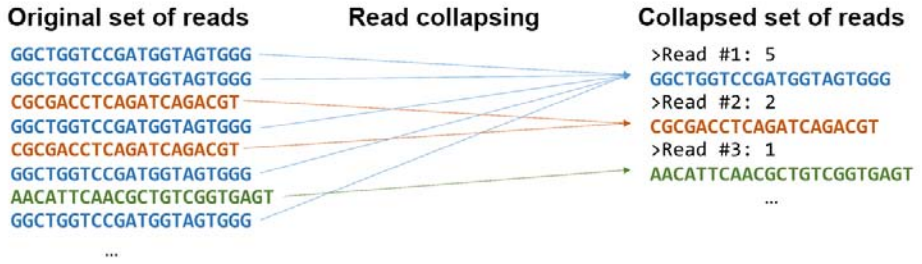


Figure 8: miRNA-seq preprocessing: Read collapsing. Identical trimmed reads from the original file are collapsed before mapping to reduce the number of alignments and speed up the process. *Adapted from [219]*

2. **Alignment and quantification**

Once identical reads have been combined, alignment can take place. Alignment of the reads can be performed against the genome or against miRNA libraries (or, more generally, several small RNA libraries). Alignment to miRNA sequences is faster as the query space is greatly reduced and it will accurately quantify most miRNA reads in most scenarios. Genome alignment can be useful for miRNA prediction or discrimination between NTA nucleotides and longer than expected mature transcripts. Regardless of mapping to the genome or libraries, a miRNA reference has to be provided for miRNA quan-

tification. The most used miRNA references are miRBase [188] and MirGeneDB [189], a more conservative database that only lists highly confident and properly annotated miRNAs.

A short read aligner such as bowtie is appropriate to deal with small RNA read mapping. The alignment should allow for sequence variants, extensions of the canonical sequence, non-templated nucleotide additions and other length variants as they have been described as biological manifestations of the miRNA. The algorithm will try to retrieve all genomic locations or miRNA entries where the read matches. Combining mapping information from the collapsed file we can get the read count for each miRNA detected in the sample.

Read count can be useful for addressing a number of questions relative to individual samples such as which miRNAs are most highly expressed. However, direct comparison among samples cannot be performed from this data as different sequencing runs may yield a different amount of reads. Different normalization approaches are available for RNA-seq data but they normally take into account the length of the mapping transcript, a negligible factor in the case of miRNAs. In the case of small RNAs, Reads per Million (RPM) seems to be appropriate.

3. **Downstream analysis**

After quantification has been completed several downstream analyses can be performed, the most common being differential expression analysis which allows to statistically determine if changes in expression or abundance of a given miRNA are greater than expected by

random variation. In the context of liquid biopsy, it can be used to discover biomarkers that are more or less abundant in patient samples compared to controls.

Another application is miRNA discovery. Reads that are mapped to the genome can form a distinct pattern typical of miRNAs. Machine learning models can then be used to assess how likely a candidate is to be a true novel miRNA [163, 220]. Further confirmation can include conservation analysis in evolutionarily close species.

Other downstream analyses include miRNA target prediction [221] and functional analysis of miRNA targets [222].

1.3.5 Other small RNAs detected by miRNA-seq

The size selection step in miRNA-seq is designed to enrich the fraction of miRNAs that get sequenced. Nevertheless, this approach can only go so far as other ncRNA of similar size or degradation products from longer transcripts will also be co-sequenced. With the possible exception of tRNA fragments, these molecules have received much less attention. Here, a brief summary containing the most relevant small RNA types is provided.

Table 1: Summary of small non-coding RNA classes detectable by miRNA-seq. *Adapted from [223]*

Class	Size	Function
Messenger RNA (mRNA)	2–5 kb	RNA that contains a coding region that directs synthesis of a protein product. Fragments can be detected

PIWI-associated RNA (piRNA)	27 nt.	RNA that directs the modification of chromatin to repress transcription.
Transfer RNA (yRNA)	100 nt.	RNA adaptor connecting an mRNA codon and the activated form of the cognate amino acid during protein synthesis on the ribosome. Normally only halves or fragments are sequenced.
Ribosomal RNA (rRNA)	120, 160, 1,868, 5,025 nt in human	RNA component of the small or large ribosomal subunit; the largest is a ribozyme. Normally present in short fragments (degradation products)
Small interfering RNA (siRNA)	22 nt.	Product of Dicer cleavage of dsRNA; when complexed with an AGO protein, induces cleavage of a perfectly-complementary target RNA
Long non-coding RNA (lncRNA)	>200 nt	Long elements transcribed by RNA polymerase II. Fragments can be detected
Small nuclear RNA (snRNA)	100-300 nt.	RNA localized in the eukaryotic cell nucleus

Small nucleolar RNA (snoRNA)	70 nt.	Essential for pre-rRNA processing or modification by serving as a guide RNA to direct a bound enzyme to either 2'-O-methylate or pseudouridylate a complementary sequence in rRNA. Unless the protocol is adapted to select longer reads, only fragments are detected.
Y RNA	∼100 nt.	RNAs that bind the Ro60 and La proteins to form the Ro ribonucleoprotein complex that mediates, between other functions, in DNA replication.
Vault RNA (vRNA)	80-150 nt	Small untranslated RNA molecules that make part of the vault complex

1.3.6 miRNA isoforms: a potential new layer of information

Until not so long ago it was believed that each pre-miRNA hairpin could generate two mature miRNAs: the 5' and 3' arms. With the arrival of NGS techniques, researchers started detecting several miRNA sequences that slightly differed from the canonical mature miRNA available in references [224] but their biological relevance remained controversial [225]. Said sequences varied in length (shorter or longer 3' or 5' end), sequence composition (different nucleotides in the read detected) or a combination of

both [226] (Figure 9). Nevertheless, some studies shown that isomiRs are not randomly distributed [225, 227] so much so that they could be used to discriminate amongst cancer types [228]. These findings suggested that the biogenesis of isomiRs might be regulated and therefore be functional. In accordance with these assumptions, researchers reported that 5' isomiR of miR-9 changed targets compared to its canonical version [227].

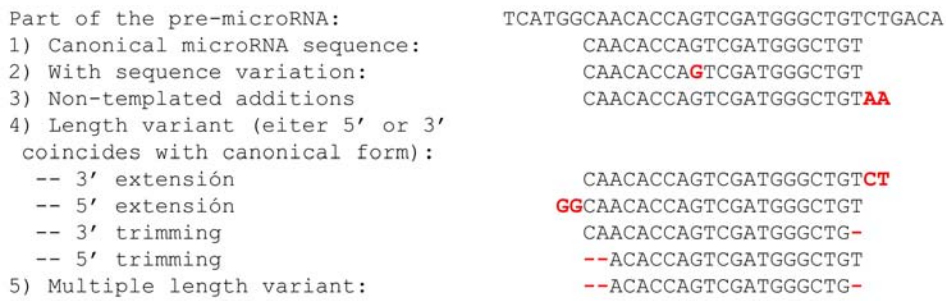


Figure 9: Classification of isomiRs (miRNA variants) ***Adapted from [219]***

Since then, the accumulated evidence strongly suggests that isomiRs have unique molecular roles and that they can target different mRNAs than their corresponding canonical miRNA both from shifting by 5' alternative cleavage [229] and from 3' uridylation [230]. Together with their cell and tissue specificity, this makes them good potential biomarkers of cancer and other diseases.

Chapter 2

Objectives

The aim of this Doctoral Thesis is to stablish a computational methodology to accurately detect and quantify potential small RNA biomarkers from sequencing data, particularly in the context of liquid biopsy of cancer patients. To this end, the following specific objectives will be addressed:

1. Improve and update *sRNAbench* to include more recent genome assemblies, bacterial and viral sequences, miRNA references and other small RNA databases. Implement preprocessing of new library preparation protocols and automated protocol detection.

2. Develop an automated workflow to acquire and profile small RNA sequencing samples from publicly available repositories.

3. Collect all publicly available miRNA-seq samples to build a reference corpus of samples for reference and reuse of the data.

4. Manually curate and organize liquid biopsy miRNAseq samples into a publicly available database that allows search of samples, studies and comparisons.

5. Develop a quality control pipeline for miRNAseq experiments to provide relative assessment of quality parameters and features using the reference corpus of miRNA-seq samples.

6. Apply the methodologies developed to miRNAseq data from a liquid biopsy study of melanoma patients

Chapter 3

Material and methods

Since this thesis is presented as a compilation of articles, the specific methodology is described in each corresponding chapter. In this section, I will present or expand methods that are cross-sectional or that were originally kept out of the articles included here because of word count limits or other constraints.

3.1 Webserver Implementation

Three interactive web sites are described as part of this thesis: *sRNAbench* [218], *liqDB* [95] and *mirnaQC* [231]. Each of these services was implemented and deployed in an independent containerized environment using Docker 20.10.8 and run on an Ubuntu 18.06.6 machine. An Apache server (version 2.4.29) redirects URL traffic from each site's subdomain to the corresponding service by means of *mod_rewrite* and port-forwarding. Web access is provided by the HTTPS protocol.

3.1.1 Framework

All three services are implemented in Django (*liqDB* and *sRNAbench* use version 2.1, *mirnaQC* uses version 3.0) and Python3 (*sRNAbench* and *liqDB* use Python 3.5.4, *mirnaQC* uses Python 3.7.2). Web server models were stored in a MariaDB database (distrib. 10.4.6) run in an independent docker container and connected to the main app using Django's MySQL

engine. Each website is served from an independent Apache server configured with WSGI.

3.1.2 Back-end

Database lookups and other backend processing was performed using java classes packaged in jar files and run using OpenJDK version 11.0.11. Database queries are performed on the aforementioned MariaDB database. Additionally, *sRNAbench* jobs are run using TORQUE, a queueing and resource managing system.

3.1.3 Front-end

HTML templates are injected with pertinent variables using Django. To improve functionality and interactability several JavaScript and jQuery scripts were included. Further JavaScript libraries were used for graphing or upload including Plotly, Ajax, DataTables and Chart.js. CSS frameworks used to provide style include Bootstrap and Font Awesome.

3.2 Automated miRNA-seq sample acquisition workflow

To periodically update the corpus of publicly available miRNA-seq samples, a Python script was implemented to query the "sra_experiment", "sra_study" and "biosample" tables from OmicIDX API [232], a project that parses and serves public genomics repository metadata. The script

uses Google's Python implementation of BigQuery, the data warehouse where OmicIDX is stored, to retrieve and match study, experiment and sample data by means of MySQL queries. The script is regularly executed using cron jobs scheduled to happen every month.

Because there is little or no active validation from SRA on experiment annotation, it is relatively frequent that some miRNA-seq experiments are misclassified by authors as ncRNA-seq or simply RNA-seq, even though they actually follow a small RNA-seq protocol. For this reason, experiments annotated as any of those three library strategies were kept for potential downstream processing after appropriate check-ups.

3.3 Automated sample processing and collection into a database

SRA experiments were downloaded using fasterq-dump from SRA-tools [233]. Once a file was downloaded, to decide whether the sequencing experiment was effectively miRNAseq , the presence of at least 20 different mature miRNA sequences was required. Samples sequenced with a different protocol or of insufficient quality were discarded. After the sequencing strategy was confirmed, a library preparation protocol detection algorithm was applied, the following were considered: TruSeq RNA Library Prep Kit v1 (Illumina), TruSeq RNA Library Prep Kit v2 (Illumina), NEBNext Small RNA Library Prep Kit (New England Biolabs), NEXTFLEX Small RNA-Seq Kit v3 (PerkinElmer), QIAseq miRNA Library Kit (QIAGEN) and 4N random adapters.

Each sample was preprocessed with *sRNAbench* using the appropriate configuration for the previously detected protocol. Quality-checked adapter-trimmed reads were collapsed and stored as fasta files. Reads were aligned using *sRNAbench* and the following mapping libraries: miRNA (miRBase release 22.1, MirGeneDB 2.0), tRNA (GtRNAdb 2.0), yRNA (from RNAcentral), vRNA (from RNAcentral), remaining non-coding RNAs (RNAcentral and ncRNA from Ensembl), mRNA (cDNA from Ensembl) and a collection of vertebrate viral genomes (retrieved from NCBI).

A MariaDB database was implemented to store count data from all libraries plus the metadata retrieved from SRA. Collapsed reads were also stored for reanalysis.

3.4 Read count adjustments and normalizations

Converting mapped reads into read count would appear as a straightforward problem: each library sequence receiving a mapping should get one count added. That's indeed a possible solution, most frequently referred to simply as read count (RC files).

Nevertheless, microRNAs are frequently members of broader families that contain several genes with highly similar mature sequences. Given also the short nature of miRNA-seq reads, multiple-mapping reads are to be expected. To deal with this, *sRNAbench* implements a solution to account for this fact without discarding multiple-mapping reads. Adjusted expression values are recalculated in the following way: each multi-mapping count is divided by the number of reference sequences to which they map.

This is termed multiple-mapping adjusted read count (RCadj files).

Another issue affects inter-sample data interpretation since different experiments, even if performed in the same sequencing run, will yield varying number of reads, i.e. read abundance is a relative variable. In practice this translates into unfair comparisons if counts are not scaled or normalized, as 1000 RC (read count) could be a lot for an experiment of 100.000 reads but very little in a different experiment of 20 million reads. In *sRNAbench*, this is solved by scaling all reads to 1 million, rendering samples comparable. Reads per Million files are normally referred to as RPM. Please note that RPKM or FPKM normalizations are not required for small RNA sequencing experiments since transcript length range is negligible. Other normalizations, like the one provided by *edgeR*, are also implemented.

Additionally, miRNA count data generated by sequencing platforms is a composition of all the sequences in the file, which are only a random portion of all the RNA actually present in the sample. This means that data can only be interpreted in relative terms. Furthermore, absolute changes in one particular RNA sequence or fraction in the sample will necessarily translate into relative changes for the rest of them. For instance, we could be comparing two samples, A and B, where the concentration of hsa-miR-21-5p is the same in both. However, sample B has an increased amount of Y RNA, which gets co-sequenced in small RNA-seq. Since the Y RNA is "competing" to get sequenced, fewer hsa-miR-21-5p reads will make it to the final count in sample B compared to A, even though both samples effectively have the same concentration. Other artifacts can also affect the expression in a similar way, such as an increased number of unmapped reads. This

problem can potentially be mitigated by using library-normalized RPM. Files using this normalization are denoted with RPMlib.

3.5 Differential expression analysis

The purpose of differential expression analysis (DE or DEA) is to statistically test whether genes, or non-coding transcripts, display significantly different expression levels between two groups or conditions (i.e. whether the difference is greater than expected by random variation). This is useful to discriminate between likely true and spurious changes. It should be kept in mind that both the effect size and the significance, measured as fold change and p-value respectively, are necessary to fully understand the impact of a reported change.

Differential expression methods normally consist on some normalization approach and a statistical test, both of which normally rely on assumptions about the distribution of the expression or abundance of the counts. Here I will briefly describe the differential expression methods that are used as part of this thesis.

3.5.1 Student's t-test

Student's t-test [234] is one of the most used statistical tests. It is performed to determine if the means of two groups are equal or not. This test assumes that both samples come from a normally distributed variable and that variances are equal. These assumptions probably don't hold for miRNA cellular expression but they might for circulating RNA. Normalized read

counts (RPM) are used as input to avoid unfair comparisons caused by different library size.

The test was carried out using the Commons Math java library, homoscedasticTTest, a two-sample two-tail t-test that compares the means of two input groups assuming homoscedasticity (i.e. both groups have the same variance).

3.5.2 *edgeR*

edgeR is a Bioconductor package that can mainly be used for differential expression analysis of sequencing data [235]. Read counts and sample/-condition annotation are to be provided as input. *EdgeR* can deal with complex study designs that take into account multiple factors to exclude technical variation.

The scaling method used by *edgeR* is the trimmed mean of M-values normalization method (TMM) [236]. This approach assumes that the majority of genes are not differentially expressed. To summarize the observed M values, both ends of the distribution are trimmed and then weighed using the inverse of the variance to account for larger variation at higher expression levels [236]. Weighed values are then modeled using a negative binomial distribution.

3.5.3 *DESeq*

DESeq works in a similar way to *edgeR*, it also requires a count matrix and sample annotations as input. *DESeq* package normalizes the counts

using size factors and models the data by means of a negative binomial [237]. However, the algorithms and models still differ and so do the results: *edgeR* has more false positives on low count but less false positives on high counts [237].

DESeq was improved on a second version which better estimates dispersion on low-count genes. Nevertheless, we kept *DESeq* as part of *sRNAtoolbox* so users can replicate previous results or run methods used in previous publications for reproducibility or comparison.

3.5.4 *DESeq2*

A newer version of *DESeq* was published in 2014 [238]. The current algorithm differs from the previous version in the method to estimate the dispersion: *DESeq* used the maximum of the fitted curve and the gene-wise dispersion estimate as the final estimate, which it tended to overestimate [238], whereas *DESeq2* sequentially estimates a prior distribution for the true dispersion values, and then provide the maximum a posteriori as final estimate [238].

This latter approach is more powerful in general terms as it provides recall rates similar to those reported by *edgeR*. However, this increased statistical power comes at the cost of less precision [239] (i.e. more false positives).

3.5.5 *NOISeq*

NOISeq largely differs from previously described methods because it's data-adaptive and non-parametric, which was intended to deal with a raise in false positives with increasing depth, a pitfall observed in other methods [240]. *NOISeq* is more effective in controlling false positives and very useful to account for low amount of replicates as well as genes at the low expression range. It's the only method here that can't be assessed through a p-value since it provides a DE probability instead.

Chapter 4

sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression

Based on the manuscript published in Nucleic Acids Research, 2019 Jul 2; 47(W1):W530-W535. doi:10.1093/nar/gkz415

Since the original publication of sRNAtoolbox in 2015, small RNA research experienced notable advances in different directions. New protocols for small RNA sequencing have become available to address important issues such as adapter ligation bias, PCR amplification artefacts or to include internal controls such as spike-in sequences. New microRNA reference databases were developed with different foci, either prioritizing accuracy (low number of false positives) or completeness (low number of false negatives). Additionally, other small RNA molecules as well as microRNA sequence and length variants (isomiRs) have continued to gain importance. Finally, the number of microRNA sequencing studies deposited in GEO nearly triplicated from 2014 (280) to 2018 (764). These developments imply that fast and easy-to-use tools for expression profiling and subsequent downstream analysis of miRNAseq data are essential to many researchers. Key features in this sRNAtoolbox release include addition of all major RNA library preparation protocols to sRNAbench and improvements in sRNAde, a tool that summarizes several aspects of small RNA sequencing studies including the detection of consensus differential expression. A special emphasis was put on the user-friendliness of the tools, for instance sRNAbench now supports parallel launching of several jobs to improve reproducibility and user time efficiency.

4.1 INTRODUCTION

Small RNA profiling by means of miRNA-seq (or small RNA-seq) is a key step in many study designs because it often precedes further downstream analysis such as screening, prediction, identification and validation of miRNAtargets or biomarker detection [241, 242]. Many different tools areavailable for the analysis of small RNA high-throughput sequencing data such as miRDeep2 [243], miRge 2.0 [220], Short-Stack [244], SeqBuster [245], sRNAbench [246] and miRTrace [247] which implements a new approach to quality control. Generally, the tools focus on certain aspects of small RNAs and are not integrated into independent pipelines for downstream analysis. In 2015, we introduced sRNAtoolbox [248], a collection of small RNA research tools built around sRNAbench, providing different downstream analysis including consensus differential expression, target prediction and analysis of unmapped reads by means of blast searches against general nucleotide databases.

The last few years have witnessed a further drop in sequencing cost that together with the advent of highly specialized service providers makes the generation of this kind of data accessible to a larger number of research groups. The increase in sequencing volume has been accompanied by the publication of new library preparation protocols, each of which involves specific pre-processing steps in the bioinformatics analysis. However, not all research groups can count on specialized staff or bioinformatics equipment, which is why flexible and user-friendly tools for small RNA research became even more valuable over the last years. Here, we present the lat-

est version of sRNAtoolbox, featuring key additions to sRNAbench and sRNAde. Apart from customizable preprocessing, sRNAbench now implements automatic processing of the five most used library preparation protocols including UMI-based (Unique Molecular Identifier) protocols and the detection of putative sequence variants. The scope was notably increased by including new reference genomes from Ensembl (release 91), bacteria and virus collections from NCBI and microRNA reference sequences from MirGeneDB. Additionally, in order to improve reproducibility and ease of use, a batch mode was developed to allow profiling of several samples at once using the same set of parameters. As for sRNAde, now consensus results for five differential expression methods are calculated together with improved visualizations of several quality and mapping statistics.

4.2 WHAT'S NEW?

Since sRNAtoolbox web-server has previously been described [248], we briefly present main novelties and changes in this section. More detailed descriptions can be found in the Data and methods section.

- **sRNAbench batch mode:** users can now provide an unlimited number of reads files through upload, URLs or SRA Run accessions. In this way, parameters only need to be specified once and are applied to all input data.
- **Reanalysis of provided files:** All provided files can be reanalysed without reuploading to the server.
- **New sRNAbench features**: Optional quality control of fastq input,

detection of sequence variants, direct availability of 6 different library preparation protocols, UMI (Unique Molecular Identifier) protocols are supported, isomiR classification can be made hierarchical (each read belongs to only one category) or fuzzy (each read can belong to several categories), input format is automatically detected to prevent inconsistent file extensions and improved feedback so most frequent input errors can be corrected by the user.

- **Visualization of genome mapped reads:** The jBrowse instance to visualize the genome mappings was replaced by links to UCSC Genome Browser or Ensembl track hubs. Additionally, direct downloads to bedGraph, big-Wig and bed files are provided so they can be analysed using specialized software like the Integrative Genome Viewer [249].

- **Differential expression:** We added two additional methods to detect differentially expressed microRNAs: a Student's t-test and DESeq2 [238] for a total of 5 different methods. Each method has its own output page which includes interactive heatmaps [250], box-plots and volcano plots to visualize differences in expression values between two groups. The consensus differentially expressed microRNAs are visualized by means of UpsetR [251], an alternative to Venn diagrams. By default, adjusted read counts (to address multiple mapping) are used to generate the expression matrixes, but matrixes for other multiple mapping methods can be found in the downloadable results.

- **Consensus target detection:** The original miRconstarget was split into two, one tool for animals and one for plants. A simple seed

detection method several folds faster than the other three (miranda, PITA and Target-Spy) was added to the animal tool.

- **Scope:** Genome sequences and annotations are automatically derived from Ensembl [252]. Current version of sRNAtoolbox contains 90 genome assemblies and several virus and bacteria collections obtained from NCBI [253].

- **Reference sequences:** microRNAs for all species included in miRBase [188] or MirGeneDB [254] can be profiled regardless of genome availability.

- **liqDB:** sRNAbench is now connected to liqDB, a small RNA database for liquid biopsy studies [95], i.e. sRNAbench output can be used to compare against liqDB profiles.

4.3 DATA AND METHODS

Input data

Input files can be uploaded to our server, be provided as URLs or as SRA Run IDs [209]. For URLs or SRA run identifiers,several files can be merged together by joining them using colons (:). For example SRR2105509:SRR2105510 would merge both SRA runs into a single job. In the previous sRNAbench version, the input format was detected based on the file extension only, i.e. *.fastq for fastq format, *.fa for fasta format and *.rc for read count format. Because sRNAbench jobs could fail due to an incorrect extension, we included now an automatic detection of the

input format to prevent those errors. Automatic detection of most common separators in read-count encoded fasta files has also been implemented.

Quality control

Two quality filters have been implemented in sRNAbench for fastq files. The 'mean' method calculates the average PhredScore of the adapter-trimmed read, filtering out those below a certain threshold. The 'min' method is stricter as it sorts out any read with at least one position below the provided threshold.

MicroRNA profiling, genome and library mode

Expression values can be obtained either using genome or library mode. In genome mode, reads are first mapped to the corresponding assembly and genome annotations of the reference sequences are used to obtain the expression values. In library mode, reads are mapped directly against the reference sequences. Both methods are described in detail in the original sRNAbench paper [246]. MicroRNA expression profiles can be obtained for all species contained in miRBase or MirGeneDB by means of the library mode. It is important to note that expression files generated with sRNAbench will list all copies of a microRNA, and therefore the name of a mature microRNA can appear several times. However in an additional column we specify the genome position or precursor name, which makes each line unique.

Two different methods are provided for multiple mapping, (i) adjusting the read count by the number of times the read maps to the genome or reference sequences and (ii) assign each read only once to the reference

sequence with the highest expression (single assignment) (see [246] for more details). The prediction of novel microRNAswas described before in the sRNAbench paper [246] and a more detailed description is available in the manual as well.

Genome mapping, bedGraph, bigWig and bed files

Adapter trimmed and quality filtered reads are mapped to the genome by means of bowtie1 [255]. By default, bowtie seed alignment is used in order to detect isomiRs (with seed length of 20 nt) and reads are only used if they have at most 10 mappings to the genome. The best mappings are retained as explained before [256]. Both parameters can be changed by the user. For the prediction of novel microRNAs, we recommend 'full read alignment' and not allowing mismatches. Some putatively interesting small RNAs like yRNAs have many copies in the genome, and therefore the maximum number of allowed mappings might need to be increased in such cases.

Reads with more mappings to the genome than specified by this threshold are not used for expression profiling but will appear as a separate category in the genome mapping plots. Those reads are labelled Highly Redundant reads and are marked with the postfix (HR).

Downloadable bedGraph files are generated summing the reads that map to a certain position. Note that in this way, each read counts fully at each position it maps (full read assignment). In the standalone version, the user can chose to adjust for multiple mappings. BedGraph files are generated irrespectively of the strand and for both strands separately (three files in total). Sometimes, it might be interesting to analyse the genome

distribution as a function of the read length (20).Therefore, we provide the bedGraph files for different length intervals: 19 nt–23 nt and all lengths for animals and 19 nt–23 nt, 24 nt and all lengths for plants given that 24 nt long reads have a very well described function in plants [257]. The bedGraph files are then converted to bigWig files using the UCSC tool bedGraphToBigWig [258]. Finally, the bedGraph files are screened and continuously mapped regions are merged together into a six-column bed file. The provided score indicates the highest expression value of the region as not all positions in a continuously mapped region will have the same expression values.

Single nucleotide variants

Single nucleotide variants (SNV) are detected based on reported mismatches. They can be due to Single Nucleotide Polymorphisms (SNPs), somatic mutations, RNA editing, sequencing or Taq polymerase errors. Therefore,when those sequence variants are analysed, strict quality control parameters should be used to control for the effect of sequencing errors and other technical artefacts. As the quality scores (Phred Scores) are not used for the detection of SNVs, this analysis can be performed for all accepted input formats. The sequence variants are detected at the level of precursor sequences, giving for each variant the precursor name, the variant type, the position, the number of mapped reads and the number of reads containing the variant.

isomiRs

The original sRNAbench version implemented only a hierarchical isomiR classification, i.e. each read is classified as only one isomiR type: canon-

ical sequence, canonical sequence with nucleotide changes, non-templated additions, 5' and 3' length variants or multiple length variants (in this hierarchical order). However, a read can have both, sequence and length variation. Therefore, we now added the possibility to explore the impact of a fuzzy classification. sRNAbench output files can be used to convert the isomiR data into standardized formats as proposed by the miRTop community (https://www.biorxiv.org/content/10.1101/505222v1, https://github.com/miRTop/mirtop).

Differential expression

The differential expression program sRNAde has undergone profound changes to provide both, an extensive summary of the whole study and the detection of consensus differential expression applying edgeR [235], DESeq [237], DESeq2 [238], NOISeq [240] and Student's t-test. Additionally, each method now has an individual page to explore the different results as well as the consensus. The output page was separated into 5 sections:

- **Results Summary:** The number of differentially over and underexpressed microRNAs per method and visualizations for the distribution of detectedRNAtypes like miRNAs, tRNAs, rRNAs etc.
- **Preprocessing/QC:** Summary of preprocessing (adapter trimmed reads, filtered reads) and read length distribution which allows to detect the presence of certain types of small RNAs (peak around 21nt corresponding to miRNAs) or artefacts like the presence of adapter dimers (reads with length 0).
- **Mapping statistics:** overview of the number of mapped and as-

signed reads.

- **miRNA and isomiR statistics:** boxplots with number of detected miRNAs, link to microRNA sequence variant analysis and isomiR statistics.

- **Differential expression:** links to the individual output pages of the five DE methods, consensus tables and its graphical representation by means of UpSet plots (equivalent to Venn diagrams).

Furthermore, sRNAde provides now three different methods to address the multiple mapping problem: (i) fullread count assignment (the full read count is assigned to all reference sequences or genome positions), (ii) adjusted read counts (divide the read count by the number of mappings) and (iii) single assignment, i.e. assign the read only once to the most expressed reference sequence.

Working examples

To demonstrate the usefulness and functionality of the newly implemented features we will concentrate on the sRNAbench (batch mode) and sRNAde tools. The batch mode is a novel extension of sRNAbench which first requests the upload of the sequencing data. We strongly recommend depositing sequencing data on an accessible server and providing the URLs by means of the corresponding textbox. The sequencing data can also be uploaded through the browser or specified by means of SRA run IDs. To illustrate the analysis of data from the public SRA (Sequence Read Archive) repository, we used the SRP046046 [259] study, which can be accessed through this page: . This study has 12 different biological samples

and one run per sample. After downloading the samples annotations (Run-Info Table), they can be imported into any spreadsheet program. In this way, the column with the run names (starting with SRR) can be easily copied and pasted into the sRNAbench (batch mode) interface (see Figure 10A). After this step, the user needs to provide information regarding the species (human) and library preparation protocol (Illumina). For each sequencing data file, a separate sRNAbench job will be created. The current state of the jobs will be shown to the user on the sRNAbench (batch mode) output page (see Figure 10B). Once all jobs have finished, the results of the individual sRNAbench jobs can be used as input for sRNAde (study summarizing and differential expression tool). In order to use sRNAde, a group label needs to be assigned to each sample to indicate the condition (such as healthy, cancer, treated, etc). The output page includes a button that will take the user through this process. Note that input samples and group information can be provided in other ways through the sRNAde page. The general structure of the sRNAde output page was previously described in the 'Data and methods' section, so here we will highlight some of the newly implemented features that will help users to better interpret their data. The read length distribution of adapter-trimmed reads (in 'Preprocessing/QC' section of sRNAde output page) contains valuable information to spot possible artefacts in the library preparation. By moving the mouse cursor over the boxplots, the values of the extreme points are depicted. Figure 10C shows that in general the number of adapter-dimer reads (the adapters have ligated directly without a fragment in between) are below 20%, however one sample (BJAB exosomes, SRR1563017) shows nearly 60% of adapter-dimers, which can indicate some issues in the library

preparation like low RNA input. In general, clear peaks corresponding to the lengths of certain RNA types should be distinguishable: microRNAs should form a narrow peak around 21– 22nt and tRNAs are known to generate fragments around 18 nt and between 32 and 33 nt. If no peaks are distinguishable or if they are very smeared out, this can indicate low RNA quality (high degradation). In Figure 10C we can observe the existence of a broad peak around the length of microRNA precursor sequences or full length tRNAs. Figure 10D shows the distribution of RNA types in the study. This graphic enables the user to obtain information about the relative quantities of miRNAs or other RNA molecules like yRNA tRNA, snoRNA or rRNA. Furthermore, the dispersion of relative frequencies of a given RNA type over the different samples can be observed. For example, the percentage of microRNAs varies between 10% and 70% in this case.

Figure 10E shows the overlap of differentially expressed microRNA between the five methods and Figure 10F depicts the overlap of microRNAs with a log2 fold-change higher than 1 or lower than -1. Note that to avoid division by 0, we add the value of 1 to the expression values. This also leads to the fact that microRNAs with extremely low expression values are less likely to produce high fold-changes due to chance alone. It can be seen that the overlap using the fold-change is very high (34 out of 49). Notice that the miRNA fold-change only depends on the normalized values of the read count input matrix (same for all methods). Therefore, the high overlap seems to imply that the normalization methods have a rather moderate impact on the fold-changes. On the other hand, there is only 1 out of 32 microRNA which shows statistically over-expression in all five methods mainly because Student's t-test and NOISeq seem to be much stricter.

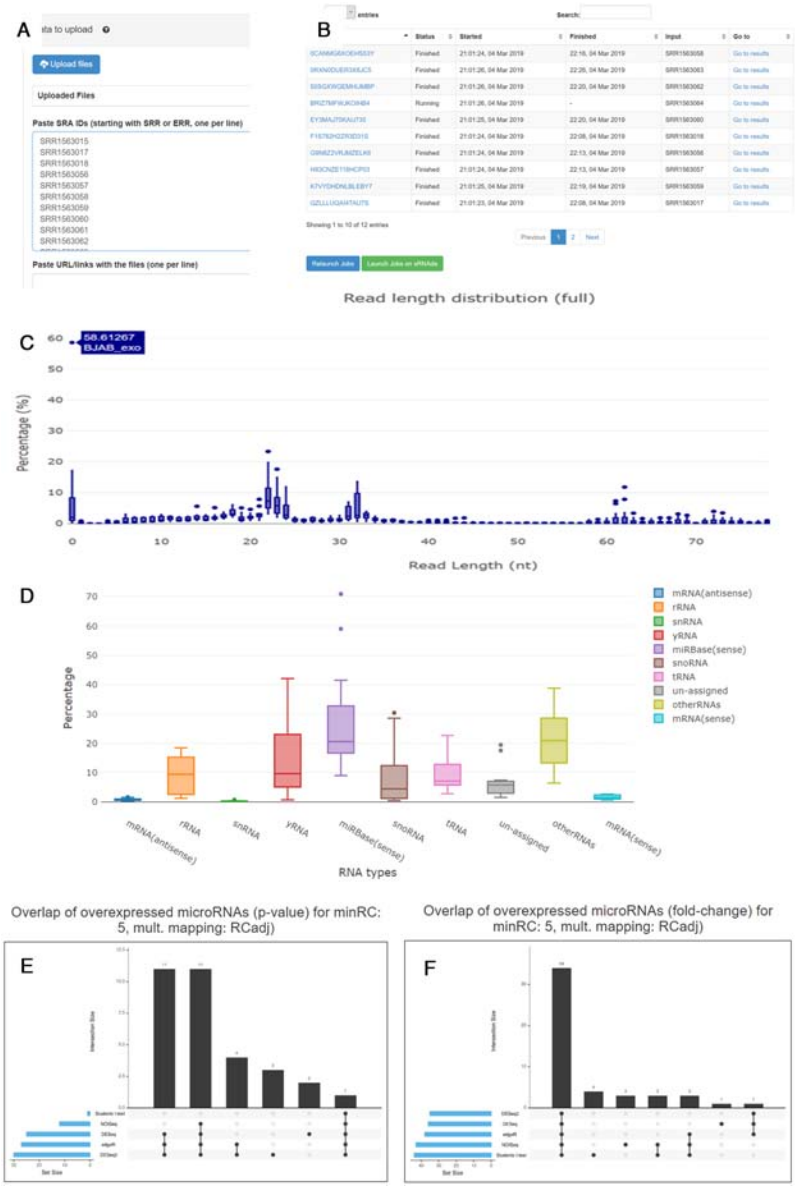


Figure 10: (**A** and **B**) The interface of the sRNAbench batch mode module and the primary result table, (**C**) The read length distribution as box-plot, i.e. the distribution of read fraction as a function of read length, (**D**) the distribution of different RNA types in the study, (**E**) the intersection of up-regulated microRNAs between the different methods and (**F**) the intersection of microRNAs with higher fold-changes than 2.

DEseq, DESeq2 and edgeR are the methods with the highest overlap (11 out of 32). This shows that the way the P-values are calculated strongly impacts the detection of differentially expressed microRNAs.

4.4 CONCLUSIONS AND OUTLOOK

Over the last years the user feedback was crucial for the evolution of sRNAtoolbox. Several of the new features and species were included upon user request. We encourage users to send feedback of any type to continue improving this collection of small RNA research tools. Upcoming improvements include, among other features, new annotations, support for user-customizable synthetic spike-ins and improved prediction of novel microRNAs.

4.5 DATA AVAILABILITY

https://arn.ugr.es/srnatoolbox/

4.6 ACKNOWLEDGEMENTS

The authors acknowledge the usage of the computational infrastructure of the Computational Epigenomics Lab of the University of Granada.

4.7 FUNDING

European Union [765492 to M.H.]; Spanish Government [AGL2017-88702-C2-2-R to M.H., J.L.O.]; Instituto de Salud Carlos III, FEDER funds [PIE16/00045 to J.A.M.]; Chair 'Doctors Galera-Requena in cancer stem cell research' to JMA and by the Ministry of Education of Spain [FPU13/05662 to R.L., IFI16/00041 to E.A.]; Strategic Research Area (SFO) program of the Swedish Research Council (to V.R.) through Stockholm University (to B.F.). Funding for open access charge: Spanish Government [AGL2017-88702-C2-2-R].

Chapter 5

liqDB: a small-RNAseq knowledge discovery database for liquid biopsy studies

Based on the manuscript published in Nucleic Acids Research, 2019 Jan 8; 47(D1):D113-D120. doi:10.1093/nar/gky981

MiRNAs are important regulators of gene expression and are frequently deregulated under pathologic conditions. They are highly stable in bodily fluids which makes them feasible candidates to become minimally invasive biomarkers. In fact, several studies already proposed circulating miRNA-based biomarkers for different types of neoplastic, cardiovascular and degenerative diseases. However, many of these studies rely on small RNA sequencing experiments that are based on different RNA extraction and processing protocols, rendering results incomparable. We generated liqDB, a database for liquid biopsy small RNA sequencing profiles that provides users with meaningful information to guide their small RNA liquid biopsy research and to overcome technical and conceptual problems. By means of a user-friendly web interface, miRNA expression profiles from 1607 manually annotated samples can be queried and explored at different levels. Result pages include downloadable expression matrices, differential expression analysis, most stably expressed miRNAs, cluster analysis and relevant visualizations by means of boxplots and heatmaps. We anticipate that liqDB will be a useful tool in liquid biopsy research as it provides a consistently annotated large compilation of experiments together with tools for reproducible analysis, comparison and hypothesis generation. LiqDB is available at http://bioinfo5.ugr.es/liqdb.

5.1 INTRODUCTION

Despite the well-established usage of blood and urine in disease detection and diagnosis, the term liquid biopsy does not appear in PubMed until 2011 in a work where breast cancer patients response to trastuzumab was monitored using circulating epithelial tumour cells (CETC) [260]. Since then, liquid biopsy has become a rapidly growing research field based on the extraction of non-solid biological material such as blood, saliva, urine or cerebrospinal fluid that can be sampled in a minimally invasive way. From this material can then be extracted, among others: protein-bound RNA molecules, vesicles such as exosomes, cell-free DNA (cfDNA), circulating tumour cells (CTC) and platelets that can be used for clinical purposes. More specifically, genotypes and methylation states of extracted DNA molecules or the abundance of RNA molecules can be screened to search for non-invasive biomarkers that allow for early diagnosis, treatment monitoring, tumour staging, relapse risk assessment and prognosis [261].

Since microRNAs were discovered in humans in 2000, their functional role as post-transcriptional repressors has been extensively studied. MicroRNA expression levels are generally altered in several pathologies including cancer [262], so they hold great potential as disease biomarkers at tissue level. Furthermore, microRNAs have been detected in virtually all bodily fluids either within exosomes [263] or bound to proteins that protect them from RNAse activity [153], both of which increase their stability and therefore their detectability. If the release of most microRNAs is fairly random [259], the assumption is that intracellular changes can be detected

in the different biofluids as well, which allows for potential applicability as diagnostic, prognostic and predictive biomarkers. In fact, several studies have already used this approach to propose miRNA-based biomarkers for different types of neoplastic [264,265], cardiovascular [266] and degenerative disease [267]. Many of these studies rely on small RNA sequencing experiments but differences in sample collection, extraction, storage, processing, library preparation and sequencing method can have a strong impact on the abundance of detected miRNAs [268]. Highly parametrized computational tools used for data analysis are yet another source of fluctuation in the obtained expression values. Note that most of these issues are not inherent to sequencing approaches, as other methods such as qRT-PCR are also affected by this panoply of possible confounding variables. Furthermore, no endogenous small RNA has been established to normalize abundance in plasma, although synthetic spike-in molecules have been proposed to address this problem [269].

In order to help to overcome the problems described above we developed liqDB, a database for small RNA expression profiles in bodily fluids. Unlike other literature-based resources such as miRandola [270] or ExoCarta [271], liqDB contains small RNA expression profiles of 1607 manually annotated samples from SRA, generated by means of a reproducible bioinformatics protocol. The database can be queried in different ways exploring 19 different biofluids or the impact of six variables like health state or RNA extraction method. Users can perform customised queries or compare uploaded data to sample sets from the database. Most important results are: downloadable expression matrixes, differentially expressed microRNAs and most stably expressed microRNAs. Visualisation of the output includes

interactive RNA distribution boxplots and heatmaps [250]. A strong increase in liquid biopsy small RNA research is to be expected over the next years and we are confident that liqDB can play a central role in organising, classifying and offering this information to researchers in the field.

5.2 SCOPE AND WEB INTERFACE

Scope

Currently, the database contains a total of 1607 coherently annotated sRNA-seq samples from 30 publically available SRA studies corresponding to 19 different biofluids. The sample annotations provided by the original authors were manually curated and unified to include relevant variables such as biofluid, gender, health state (healthy/not healthy), RNA extraction protocol, exosome isolation (yes or no) and RNA library preparation protocol, all of which can have an impact on the miRNA abundance acting as confounding variables for the variable of interest, usually related to the health state.

The database can be used and queried in five different ways:

- Browse studies: For each of the 30 SRA projects in the database, results were pre-calculated. Differential expression is calculated whenever possible for the relevant variables (mostly health state and gender although RNA extraction, library preparation methods and different biofluids were analysed for some studies). Additionally to the miRNA expression profiles, other results are generated (see Result Page below for more details). Frequently, users are interested in a particular

project focused on a specific cancer or biofluid and this information can be quickly accessed in this way.

- Search samples: Users can customize a set of samples by selecting query values for 6 variables and a threshold number of miRNA-mapping reads in order to improve confidence in the results. The *Basic statistics* section can help users find the most adequate threshold for their desired query. At a second stage, the preselected set of samples can be manually refined. Most highly expressed, most fluctuating and most stably expressed microRNAs can be obtained, among other features. The downloadable expression matrix can be used for further downstream analysis (see Result Page below for more details).

- Search miRNAs: Users interested in one particular miRNA can search it and analyse its expression values as a function of the different variables.

- Compare two datasets: Similarly to the selection of one set, the user can define two different sets of samples. Additionally to the general output results, differentially expressed microRNAs between the two sets will be calculated.

- Compare with user data: This tool allows comparison of a selected set of samples from the database to user-provided samples. Users should first profile their sequencing reads input file using sRNAbench from the sRNAtoolbox server [248]. Subsequently, the job IDs can be used as input as well as relevant query variables to generate a standard result page including differential expression between database and uploaded samples.

Results pages

The standard results page includes the sections described below.

- Overview/query results: The full annotation of all selected samples.
- miRNA profiles: a sortable and searchable expression matrix with adjusted RPM values (Reads Per Million), a boxplot of the 20 most abundant microRNAs, a pie-chart showing the relative frequency of the 10 most abundant microRNAs, the 20 microRNAs with highest coefficient of variation (CV) and the 20 microRNAs with lowest CV. The CV is calculated as the standard deviation of the adjusted RPM expression values divided by the mean value. It is a standardized measure of the dispersion of RPM values which does not depend on the magnitude and therefore allows to compare the dispersion of highly and lowly abundant microRNAs. The expression of microRNAs with lowest CV is less affected by the different variables in the analysis and could be used as reference microRNAs to normalize or standardize qPCR validation experiments.
- sRNA types distribution: Proportions of reads assigned to each of the different small RNA types (miRNA, tRNA, yRNA, ribosomal RNA, etc) are shown in a table and the 10 most frequent categories are depicted in boxplots. The category 'Un-assigned' contains genome mapped reads which could not be assigned to any annotation.
- Species Distribution: A high number of reads that cannot be mapped to the genome can indicate contamination or the presence of genetic material from symbionts or parasites. In order to address this ques-

tion, liqDB summarizes the mapping to the human genome, virus and bacteria collections showing the relative frequencies of reads assigned to the different species. If a read maps to more than 20 loci in the genome, it will not be used for expression profiling. Those reads get the label 'HR' for highly redundant. hsa-HR will therefore refer to the relative frequency of reads that map more than 20 times to the human genome. Furthermore, a read can map with the same quality (number of mismatches and length) to different indexes. In this cases, a new category is generated mentioning all genomes separated by '-'. For example human-virus-hsawill refer to the number and percentage of reads that map both to the human genome and the virus collection.

- Download: Expression matrices are available for download as well as a zip file with the complete analysis.

- Differential Expression: if available, pre-calculated study specific comparisons or generated from user-provided groups are displayed. Differentially expressed microRNAs are calculated using two-sided t-test on Reads Per Million values. Subsequently, p-values are corrected for multiple testing applying the Bonferroni procedure. Boxplots of the most abundant differentially expressed microRNAs are displayed and a link to a heatmap is provided. Files for the complete analysis can be found in the download section. Note that the output for 'Compare two datasets' (user selected sets) will always contain one pairwise comparison, while the differential expression section in 'Browse studies' might contain several pairwise comparisons if more than two groups exist for one variable (e.g. SRP061240 with three

different cancer types and healthy controls).

The microRNA search generates one output page with several boxplots depicting the expression of the microRNAs as a function of the different variables.

5.3 DATA AND METHODS

Database construction

All expression values and metadata were uploaded to a MySQL database. The interactive web interface was implemented using the Django framework together with Bootstrap, javascript and the plotly package (Plotly Technologies, 2015, Collaborative data science, https://plot.ly) for interactive data visualization. A backend java program connected to the database calculates and prepares the raw results files. Apache2 was chosen as HTTP webserver.

Data collection and processing

Suitable data was searched using the NCBI based SRA (short read archive) repository and publications included in PubMed. The data was downloaded in sra format and converted with fastq-dump to standard fastq format. For expression profiling, sRNAbench, the successor program of miRanalyzer [256] was used. After removing adapter or barcode sequences, the obtained clean reads are collapsed into unique reads (UR) assigning a read count (RC) to each unique read sequence. By means of bowtie1 [253] the collapsed reads are mapped simultaneously to the human genome (GRCh38, patch

10), a collection of bacteria (Bacteria Ensembl, Release 39) as well as to human virus sequences (Human virus from EnsemblGenomes) and only the best mapping reads are retained as described before [256]. One mismatch within the 19 nt seed region was allowed in the mapping process.

The genome mapped reads are then assigned successively to several reference libraries in a hierarchical way in the exact order described below. After each step, reads can only be assigned to one library in order to avoid cross-library matches. For example, after mapping miRNA reads, those are removed and cannot be assigned to tRNAs or other small RNA annotations.

- miRNAs: miRBase v22 [272] and miRGeneDB v.2 (Fromm et al., MirGeneDB2.0: the curated microRNA Gene Database. bioRxiv, https://doi.org/10.1101/258749). We produced a merged annotation using miRBase names adding miRGeneDB sequences if those are not annotated in miRBase. Human and several virus annotations are used for profiling.
- tRNAs: GtRNAdb, a genomic tRNA database [273]
- vault-RNA, yRNA and guide RNAs extracted from RefSeq [274]
- Non-coding RNAs from Ensembl [252]
- Non-coding RNAs from RNAcentral release 9 [275]
- RNA sequences of coding genes from Ensembl (Release 91) [252]

Please note that no further filtering of the miRBase reference sequences was performed which implies that, very likely, false positive miRNAs exist

in the database. An example of this is miR-1246, which turned out to be a U2 small nuclear RNA (RNU2-1) fragment useful to discriminate tumours from controls [276].

Expression values and relative frequencies

The microRNA expression values are calculated based on raw read counts adjusted for multiple mapping. The expression matrixes in RPM or adjusted raw read counts can be downloaded. The latter is the required format for most differential expression packages like DEseq [237] and edgeR [235].

5.4 WORKING EXAMPLES

The influence of library preparation protocols on the microRNA profiles

The influence of different library preparation protocols on plasma-derived exosomal RNA frequencies was first studied in 2013 by Huang et al. [265]. These authors report that the five most abundant microRNAs account for 49% of all miRNA mapped reads. Since this study is available at liqDB, we can replicate these results by simply clicking on *Browse liqDB*, *Browse studies*, finding this study (SRP020486) and then navigating to the miRNA Profiles tab. In the miRNA profiles table it can be confirmed that liqDB lists the same five miRNAs as the most frequent to make up for 46% of miRNA mapped reads. Small differences are likely due to different miRBase versions as well as the inclusion of miRGeneDB in liqDB. Figure 12A shows the differentially expressed miRNAs detected for this study, i.e. the miRNAs which are highly influenced by the library preparation protocol.

Since there are only two Illumina samples, statistical significance can only result from Bioo Scientific (NEXTflex) vs. NEBnext protocol comparisons. Hsa-miR-128 is the most expressed microRNA that is highly affected by the protocol. Figure 12A gives a median RPM for NEBnext protocol of 237 658 (243 837 in the original article) while NEXTflex only yields 4655 (4569) RPM, nearly 2 orders of magnitude below. RPM values from liqDB and the original article are therefore nearly identical.

However, since many more samples are included in liqDB, we can extend this analysis to a larger query in order to increase confidence in the results. For instance, we can select plasma from all healthy subjects that used miRNeasy RNA extraction and compare the samples that applied Illumina library preparation protocol to those that used NEBnext. To do so, we go to Compare datasets > Compare two sets of samples and we set the selectors (see Figure 11A) for both groups following the desired parameters. We included a minimum threshold of 200,000 miRNA reads to increase the robustness of the outcome and then proceeded with all samples. While this comparison yields three out of six most abundant microRNAs in common with the SRP020486 results, it also reveals that 11 out of the 20 most abundant microRNAs are differentially expressed. This confirms again that the miRNA abundance in plasma is strongly affected by the library preparation protocol. The hierarchical clustering (see Figure 12C) shows as well that samples are grouped mostly by library protocol, although there is also one cluster with NEBnext and Illumina samples which indicates the existence of other variables that influence the miRNA expression profiles.

Recently, the influence of sample processing in miRNA sequencing was

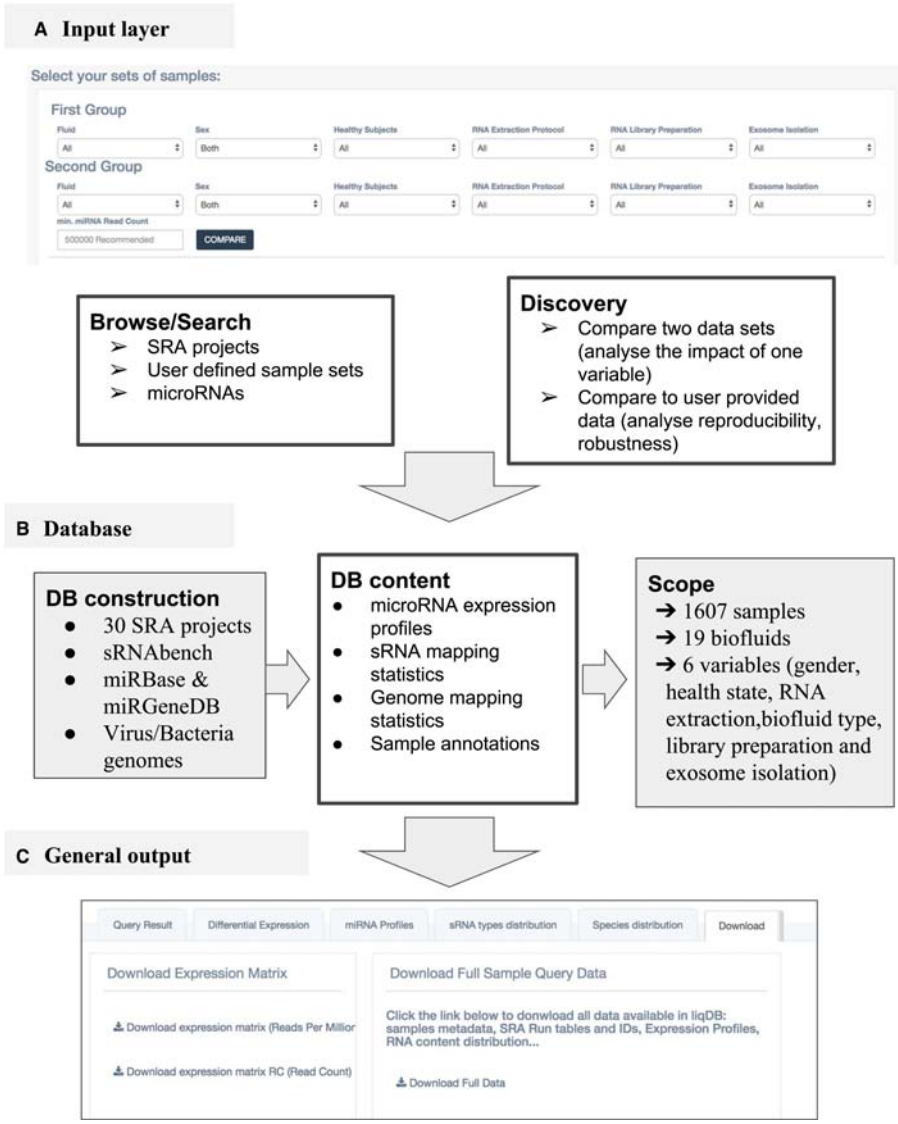


Figure 11: A schematic overview of liqDB. (A) the database can be queried in five different ways, either by browsing pre-calculated content or by instant processing of user-defined sets of samples. **(B)** liqDB was populated with 1607 samples from 19 different biofluids. The profiling of the data is carried out by means of sRNAbench (16) using both miRBase and miRGeneDB as annotations. **(C)** The general output consists of several sections, including miRNA profiles, differential expression (only if applicable) and download (shown in the figure).

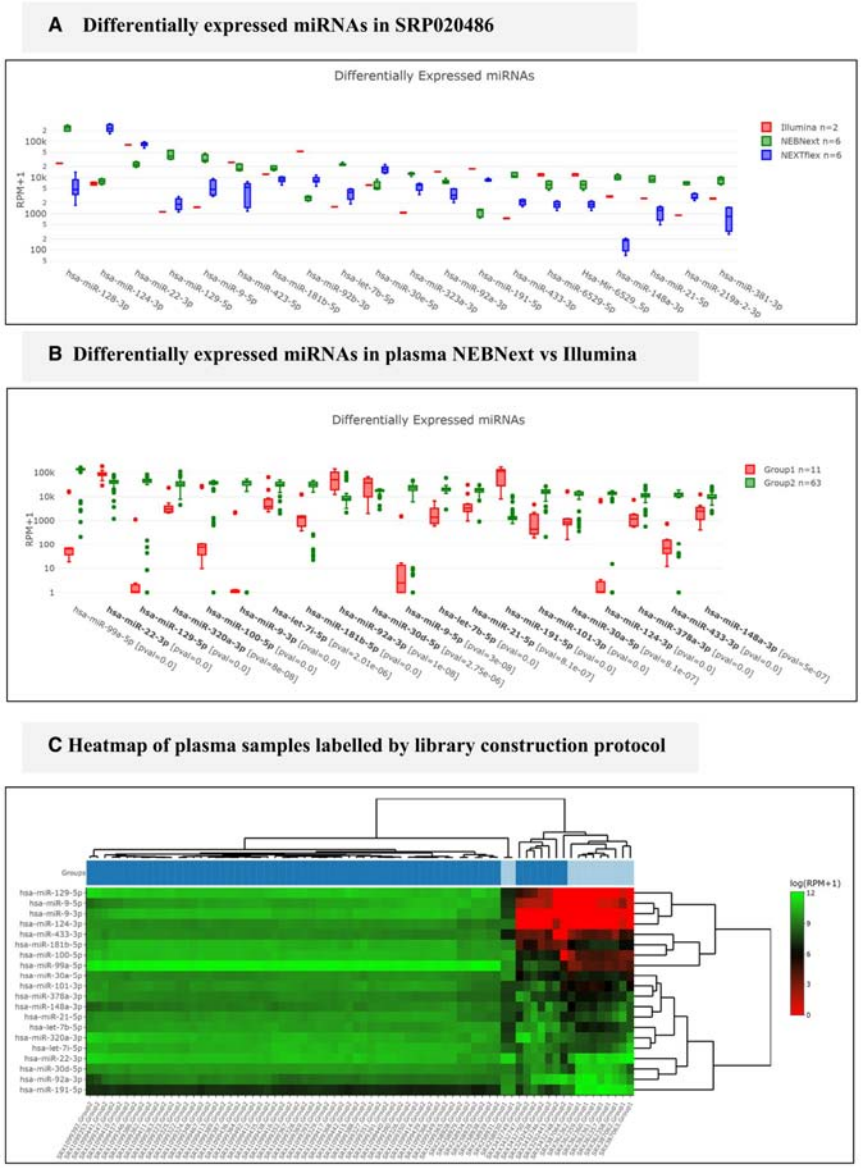


Figure 12: **(A)** Differentially expressed genes for three different library preparation protocols. **(B)** Differentially expressed miRNAs between NEBnext and Illumina protocol using all plasma samples in liqDB. (C) Most Illumina samples (bright blue) cluster together except for two of them in the middle of the NEBnext cluster (dark blue).

evaluated on a large scale study among nine different laboratories [277] confirming the strong impact of library preparation and other variables.

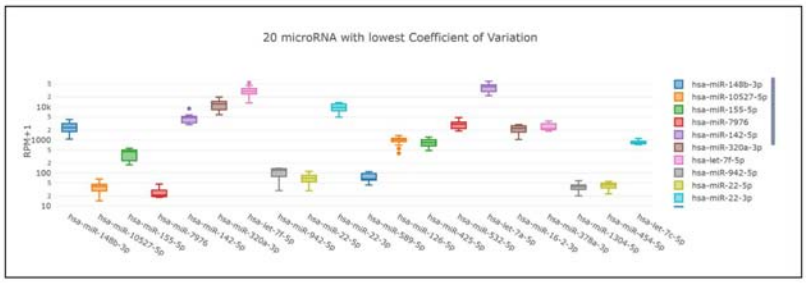
Finding least variable miRNAs in serum samples

For some downstream validation experiments such as qRT-PCR, it may be relevant to know which miRNAs are most stably expressed in order to use them as reference. In this case, we queried the database for serum samples using the following workflow: *Browse liqDB > Search samples*; then a series of selectors are displayed. For these, we chose *serum* for Fluid, *Both* for Sex, *True* for Healthy Subjects, *miRNeasy* for RNA extraction Protocol and *NEBnext* for RNA Library preparation and then clicked Filter. To avoid effect of other confounding factors, we limited the analysis to only one extraction and one library protocol. Subsequently, we kept all samples for analysis by clicking on *Proceed with all samples*. In the results page, we then navigate to the miRNA Profiles tab where the last graph will show the miRNAs with the lowest variability for the specified query (Figure 13A).

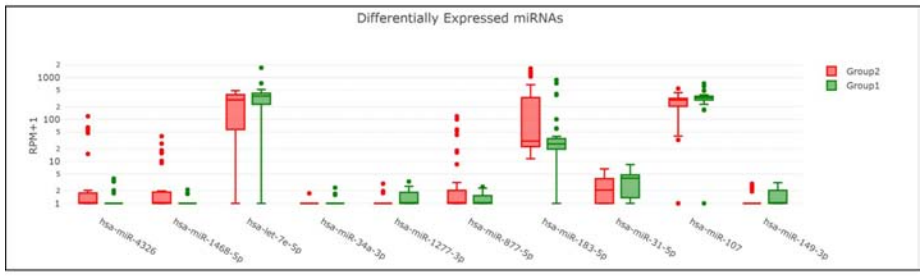
Comparing two sets of samples: plasma of men versus plasma of women

In order to compare two sets of samples in the most realistically possible way, potential confounding variables should be controlled for both sets. In this case we analysed differences in plasma between men and women. To do so, we first navigated to *Compare Datasets > Compare two sets of samples*. Similarly to the previous example, we will find two sets of selectors: one for the first group of our comparison and another for the second (See Figure

A Exploring least variable miRNAs



B Differentially expre ssed microRNAs: male vs female



C Heatmap from two user-defined sets of samples

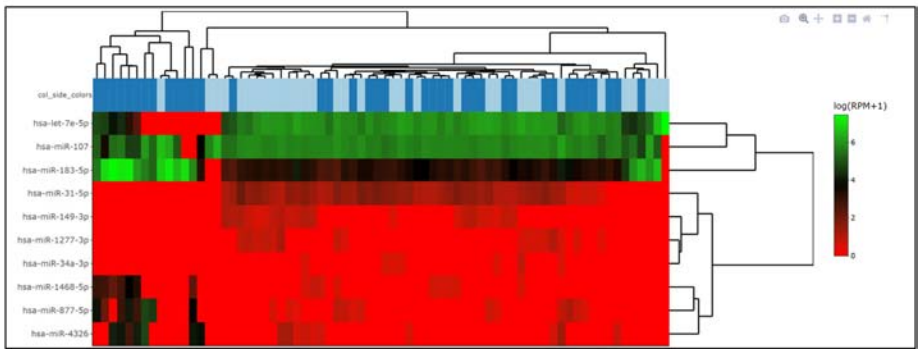


Figure 13: Examples from the web interface. (A) Boxplot of least variable miRNAs, candidates to control downstream validation. **(B)** Example of differentially expressed miRNAs boxplots. **(C)** Example of heatmap displaying clustering of plasma samples using differentially expressed miRNAs. Men are marked in dark blue and women in bright. Generated using heatmaply (15).

11A). For both selectors, we set Fluid to *plasma*, Healthy subjects to *True*, Extraction protocol to *miRNeasy* and Library protocol to *NEBnext*; for Group 1 Sex will be *male* and for Group 2 *female*. Once in the page of results, we navigated to the *Differential Expression* tab.

The boxplots graph (see Figure 13B) displays the 10 differentially expressed miRNAs with the highest expression values and then individual samples can be clustered and visualized in a heatmap (see Figure 13C).

5.5 CONCLUSIONS AND OUTLOOK

Given that miRNAs are highly stable and relatively detectable in most biofluids, it is safe to assume that they will play important roles in the fast growing field of liquid biopsy. Many projects use high-throughput sequencing approaches in the exploratory phase of biomarker discovery making the data publicly available through repositories such as SRA or GEO. In order to structure, organize and unify this vast amount of information into an interactive database, major problems like incomplete or inconsistent sample annotations need to be solved. liqDB is the first database that provides researchers from the liquid biopsy field with browse- and downloadable coherently annotated datasets generated using the same bioinformatics protocol. Furthermore the database allows the comparison to external data therefore enabling the generation and testing of new hypothesis. We also encourage researchers to share their data through SRA and submit the accession to liqDB or just to point out any overlooked SRA projects that might be suitable for inclusion.

In the short term, there are three main improvements planned. First, other small RNA derived sequences like isomiRs or fragments of tRNA, yRNA or vault-RNA molecules will be included in liqDB as they all might have biomarker potential [278]. Secondly, we will add different quality related flags to the sample annotations so the user can decide to exclude lower quality samples [279]. And finally, since differential expression is one of the key analysis for gene expression data, we will improve this feature by adding online support for DEseq [237] and EdgeR [235]. These additions will be useful for exploratory analysis of the data as well as for the analysis of confounding variables.

In summary, we anticipate that liqDB will be a useful tool for liquid biopsy researchers as it can help to develop standardized and stable protocols opening the door to reproducible reanalysis, realistic comparison and hypothesis generation, important tasks to avoid unnecessary validation of defective biomarker candidates which could prevent the discovery of actually useful biomarkers.

5.6 ACKNOWLEDGEMENTS

The authors acknowledge the usage of the computational infrastructure of the Computational Epigenomics Lab of the University of Granada.

5.7 FUNDING

European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie [grant agreement ELBA, 765492 to M.H. and D.K.L.]; Spanish Government [AGL2017-88702-C2-2-R to M.H.]; Instituto de Salud Carlos III, FEDER funds [PIE16/00045 to J.A.M.]; Chair 'Doctors Galera-Requena in cancer stem cell research' (to J.A.M.); Ministry of Education of Spain [FPU13/05662 to R.L. and IFI16/00041 to E.A.]. Funding for open access charge: Instituto de Salud Carlos III (FEDER funds) [PIE16/00045].

Chapter 6

mirnaQC: a webserver for comparative quality control of miRNA-seq data

Based on the manuscript published in Nucleic Acids Research, 2020 Jul 2; 48(W1):W262-W267 doi:10.1093/nar/gkaa452

Although miRNA-seq is extensively used in many different fields, its quality control is frequently restricted to a PhredScore-based filter. Other important quality related aspects like microRNA yield, the fraction of putative degradation products (such as rRNA fragments) or the percentage of adapter-dimers are hard to assess using absolute thresholds. Here we present mirnaQC, a webserver that relies on 34 quality parameters to assist in miRNA-seq quality control. To improve their interpretability, quality attributes are ranked using a reference distribution obtained from over 36 000 publicly available miRNA-seq datasets. Accepted input formats include FASTQ and SRA accessions. The results page contains several sections that deal with putative technical artefacts related to library preparation, sequencing, contamination or yield. Different visualisations, including PCA and heatmaps, are available to help users identify underlying issues. Finally, we show the usefulness of this approach by analysing two publicly available datasets and discussing the different quality issues that can be detected using mirnaQC.

6.1 INTRODUCTION

Different aspects of miRNA-seq such as RNA extraction, storage conditions and sample processing together with the chosen library preparation protocol have a great impact on the obtained sequencing results [280, 281]. In any bioinformatics analysis of high-throughput sequencing data, quality control (QC) is the key step to reveal the existence of technical artefacts. Neglecting this step can lead to both false discoveries and failure to identify the existing biological signal. The processing of miRNA-seq data is no exception, and QC approaches should focus on measurable sample features that can be linked to quality aspects. Moreover, whenever possible, these quality parameters should hint or point out specific technical artefacts. This approach would offer the user the chance to take appropriate actions like excluding low-quality samples from the analysis or applying statistical models in order to correct for such technical variation when possible like in the case of batch effects [282].

Several sample attributes are generally calculated in sequencing experiments including the total number of sequenced reads, number of adapter-trimmed and filtered reads, percentage of mapped/unmapped reads and PhredScore to measure the quality of the sequencing. Besides these general statistics, many pipelines such as sRNAbench [218], mirTrace [247] or miRge [220] implement measurements that are specifically useful for miRNA-seq analysis like number of unique reads, percentage of miRNA-mapping reads, read length distribution and relative abundance of fragments from other RNA types (mostly tRNA and rRNA). Many of these

parameters are clearly relevant for quality. Some indicate good quality samples when they hold high values (number of miRNA reads, total number of reads) while others (percentage of rRNA, adapter dimers) do it when they are low. Some of these features can be directly linked to a particular artefact, like a high percentage of adapter dimers which is normally caused by issues with adapters and/or input RNA concentrations [283]. Other measurements, like smeared out read length distributions, can also be attributed to specific problems, in this case RNA degradation. However, most of them can be affected by several different artefacts thus it is frequently not possible to directly reveal specific technical issues when considering each quality feature individually. For instance, the yield in microRNAs can be influenced by any artefacts that impact the total read yield including contamination.

Regardless of the values good quality measurements should take, the context-free interpretation of sample features is generally not straightforward. For example, as an obvious source of unwanted fragments, rRNA presence should be minimized, but it is difficult and arbitrary to establish a specific threshold (5%, 10%, 20%) to discard samples. Therefore, rather than working with predefined or user-provided values that are hard to justify, a more agnostic approach would rely on relative values calculated from a background of comparable experiments (i.e. similar samples) which would in turn simplify the interpretation of the QC outcome.

A vast amount of publicly available data exists that can be exploited for purposes beyond their original goal [284]. To generate a reference corpus of experiments that can be used to rank quality features, we downloaded over

36 000 raw sequencing datasets from the Sequence Read Archive (SRA) covering most model species. Samples were first processed using sRNAbench and then 34 quality features were extracted from each sample and subsequently organised into the reference set. Furthermore, sample metadata is used to tailor comparisons to more relevant sets of experiments (i.e. samples from the same species and/or processed with the same library preparation protocol).

In contrast with previously available software [220, 247], mirnaQC calculates absolute and relative values for several quality-related features for a set of miRNA-seq samples. Input data can be uploaded as FASTQ files or provided as SRA run accessions that will subsequently be ranked making use of the reference corpus mentioned above. An apparent advantage of this approach is that fixed thresholds are no longer needed and decisions can be made based on background statistics. Users can explore mirnaQC results by means of interactive plots and tables that hold both absolute and relative values of the 34 quality attributes. The output report is structured into several categories trying to relate the quality attributes to the different possible technical artefacts. This approach can help to identify low quality samples or reveal issues in the sample processing which is extremely important for protocol optimisation.

6.2 mirnaQC SAMPLE FEATURES AND QUALITY MEASURES

The success of a small RNA sequencing run depends on many different factors including RNA quality, quantity and purity, an optimized library processing protocol and the sequencing itself. However, it is not always easy or even possible to directly relate features extracted from sequencing data to any technical artefacts. mirnaQC calculates and ranks several quality parameters conceived to hint problems in the different aspects involved in the preparation of miRNA sequencing libraries. Below we describe the different sections and, wherever possible, the putative artefacts or quality issues that can be derived from them.

Sequencing yield

This section focuses on the amount of reads and the fraction that can be assigned to known miRNAs. Generally, parameters in this category (percentage of valid reads, detected microRNAs) indicate high quality when they hold high values. Low numbers (especially for the percentage of valid input reads) can be related to problems in RNA processing or low input material. Some sources like exosomes extracted from bodily fluids however, are known to hold low levels of miRNA, thus high numbers should not be expected for all sample types even for high quality libraries.

Library quality

In this category we list the number of reads that are filtered out due to minimum length (15nt), the percentage of ribosomal RNA and the per-

centage of short reads (15–17 nt). Their presence may be attributed to degradation products from longer RNA molecules as no small RNAs are known in this length range.

High percentages of adapter-dimers (0, 1 or 2 nt fragments after trimming) normally indicate issues with the ratio of adapter to input RNA concentration. In practice, it is very difficult to completely avoid adapter-dimers, especially in low input samples such as blood. Nevertheless the percentile may still be useful as it might show potential for improvement.

Ultra-short reads are defined as fragments with lengths between 3nt and 14nt (both inclusive).

Library complexity

In general it is also interesting to assess the complexity of the sample since low complexity libraries provide very little information, even for otherwise high-quality datasets. Several measurements are provided to grasp the complexity at two levels that should be interpreted together:

- Sequencing library complexity: This is calculated as the ratio of the total number of reads to unique reads. Lower values suggest higher RNA diversity but it can also be caused by degradation.
- miRNA complexity: Frequently most microRNA reads correspond to few miRNA genes preventing lowly expressed miRNAs from being detected. Several measures are given to estimate complexity at this level: (i) percentage of miRNA expression assigned to the first, the first 5 and first 20 most expressed miRNAs, (ii) the number of miRNAs required to reach 50%, 75% and 95% of the total miRNA

expression.

Putative contamination

The percentage of reads that could not be mapped to the species' genome is calculated. Contamination is subsequently estimated by mapping against a collection of bacterial and viral genomes.

Read length distribution

A narrow peak around 22 nucleotides in the read length distribution indicates good quality samples whereas degraded or poor RNA quality manifests in a broader distribution. Furthermore, it is clear that the 22nt peak should be present for miRNA assigned reads and RNA quality issues might exist if samples deviate from this.

We summarise the miRNA read length distribution in several ways: mean length, mode of the distribution, the fraction of reads with lengths 21, 22 or 23 nt, the standard deviation and the skewness of the distribution.

RNA composition

The relative abundance of other RNA molecules is automatically profiled using the sRNAtoolbox database [248]. Most of these longer RNA species (rRNA,mRNA, lincRNA) are not known to be processed into smaller molecules that can be picked up by miRNA-seq. Their presence is a symptom of RNA degradation since smaller fragments are randomly generated and then sequenced. Among these, rRNA is typically used because it's the most abundant one.

Sequencing quality

Sequencing quality is calculated by means of FastQC [285]. We determine the mean values of the different percentiles provided by the program over all positions of the read.

6.3 GENERATION OF THE miRNA-seq REFERENCE CORPUS (BACKGROUND KNOWLEDGE)

The vital part of the presented quality control tool mirnaQC is the comparison corpus of miRNA-seq data that is used to rank user's samples. Using OmicIDXR API we obtained a list of >3000 SRA studies that were annotated as 'miRNA-Seq' or 'ncRNA-Seq' (several 'RNA-Seq' were also included after checking they were in fact 'miRNA-Seq' datasets).

For each study we performed the following steps:

- Read the meta-data for a study generating one entry per experiment (SRX level)
- Download all SRR files that correspond to this SRX by means of fastq-dump (fastq.gz)
- Detect the library preparation protocol
- Analyse the small RNA sequencing data with sRNAbench using all available annotations from sRNAtoolboxDB
- Upload sRNAbench results to a MySQL database

In total we analysed 36 338 samples from 30 different species. We distinguish 8 different protocols: Illumina, Illumina-2 (3' adapter sequence), Next England Biolabs (NEBnext), Qiagen-UMI, NextFlex, adapter trimmed, SoliD and all others (custom). Over 500 billion sequencing reads were analysed.

6.4 mirnaQC WORKFLOW AND IMPLEMENTATION

An overview of the mirnaQC workflow is displayed on Figure 14. The only required input is sequencing data in FASTQ format (or SRR accessions) although sample species and library protocol information is recommended if known. If the protocol or species are not provided by the user an automatic detection algorithm, trained with a set of manually curated samples from liqDB [95], will find the right input parameters. Condition or group information can also be provided (optional). All files belonging to a given group should be compressed into a single *.zip*, *tar.gz* or *.7z* file and then separately uploaded. The file names will be used as group labels, and this information will appear in some of the plots.

Input data is subsequently processed by sRNAbench in two steps: First reads are simultaneously mapped to the species genome and a collection of virus and bacterial genomes from sRNAtoolboxDB [248] allowing one mismatch. Preference is given to the reference genome in case of multiple mapping reads. In the second step reads are mapped to microRNA reference libraries [188, 189], RNAcentral [275] and Ensembl annotations [252]

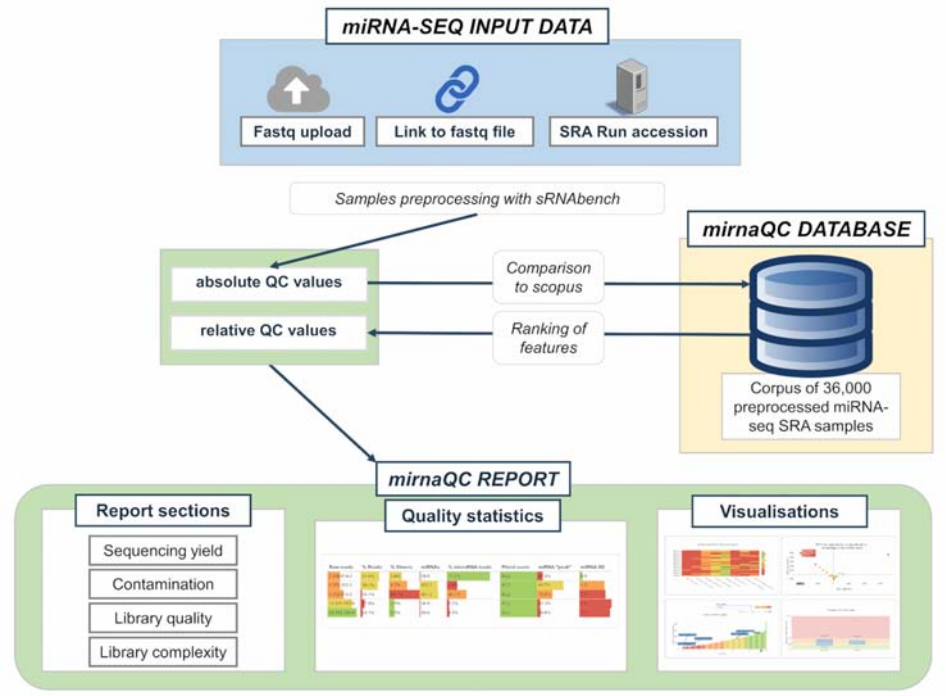


Figure 14: A schema of the front end and workflow of mirnaQC. Some features of the quality report are depicted at the bottom.

for ncRNA and mRNA. Note that although samples are mapped to both reference libraries, miRBase and MirGeneDB, currently the miRNA related figures are extracted from the miRBase mappings.

Out of both sRNAbench output folders we extract a total of 34 quality attributes that are next compared to 5 different reference sets: (i) samples from the same kingdom (animals and plants), (ii) samples from the same species, (iii) samples from the same kingdom and protocol, (iv) samples from the same species and protocol and (v) low-input samples (defined as those obtained from bodily fluids). Each comparison can be browsed separately on the output page.

The processing pipeline is a java programme that includes sRNAbench, Bowtie [255] and a MySQL client to query the reference corpus. The web interface was developed using Python and Django and runs on Apache. The results report includes the six sections described in the *mirnaQC sample features and quality measures* with tables and styles from multiQC [286] and Plotly visualisations. Both absolute values and percentiles are displayed and highlighted using a quartile colour code (see Figure 15C).

6.5 WORKING EXAMPLE

To show the usefulness of this tool, we analysed two publically available studies. Basic statistics from the first dataset, one of the earliest large studies designed to detect cervical cancer [287], can be seen on Figure 15A. To help users identify potential issues, the quality parameters and their percentiles are displayed using a quartile-based colour code (from better to worse values: green, yellow, orange and red). Using this guide, several problems can be identified: with few exceptions, most parameters rank on the third (orange, Q3) and fourth quartiles (red, Q4). More specifically, miRNA 'peak' values show that a rather low percentage of microRNA reads have lengths between 21 and 23nt in the majority of samples. This means that although those reads can be assigned to miRNA reference sequences, they do not correspond to the canonical miRNA lengths. This hints an RNA processing issue that might still be tolerated if all samples are similarly affected, which can indicate either systematic artefacts or biological reasons.

It may also happen that not all samples are equally affected by a quality issue, which can be more problematic if two or more conditions are to

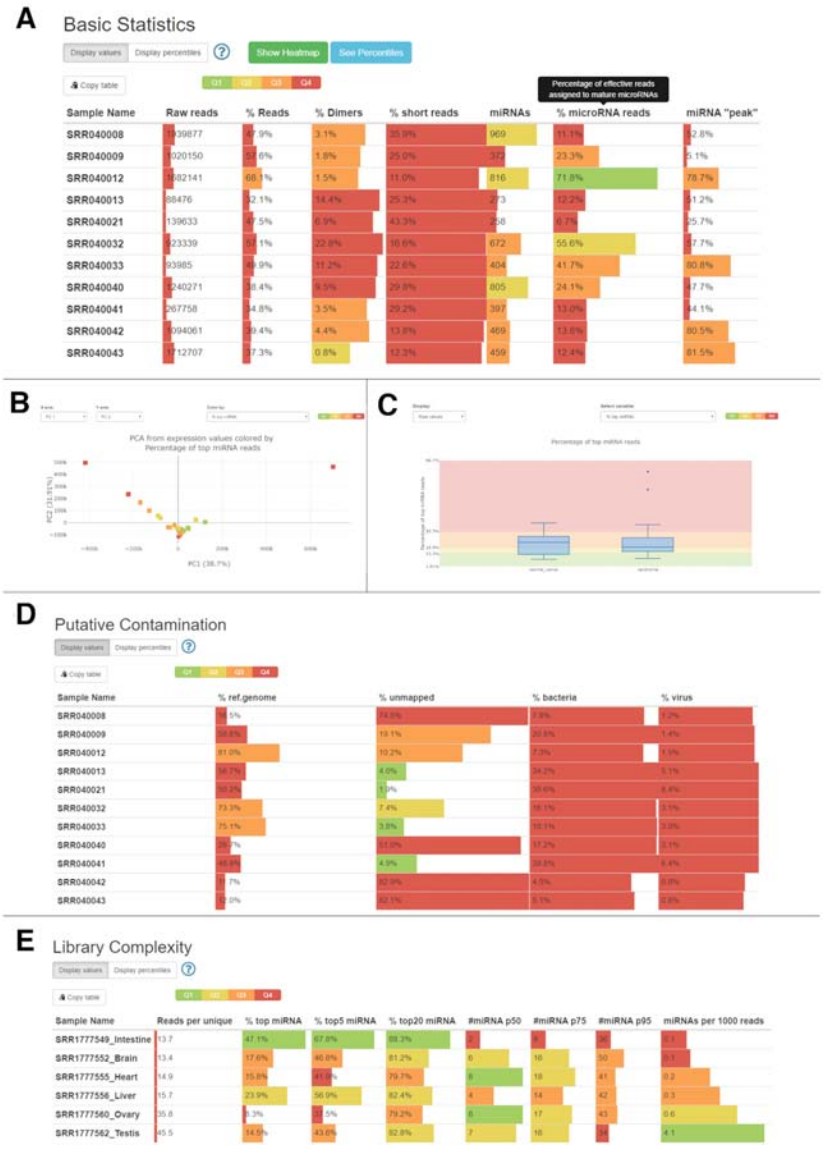


Figure 15: Different examples of mirnaQC sections and visualisations. **(A–D)** display different quality aspects of a cancer study. **(E)** shows the output of different sample complexity measures for different tissue types and two columns from Basic Statistics section at the right.

be compared. mirnaQC allows users to assign samples to conditions in order to explore this possibility. Figure 15B shows a PCA plot of the expression values of the 50 most expressed miRNAs. Users can decide which quality attribute should be used to colour the markers, in this case we used '% top miRNA' (the percentage of reads assigned to the most abundant microRNA). This graph shows that the two outlier samples are much less complex than the rest. Furthermore, because conditions are marked with different symbols (control-circles and carcinoma-squares), we know that these two samples belong to the same group. Keeping such samples in the analysis is not recommended since they will certainly bias the results. Figure 15E displays the distribution of this feature for both groups by means of boxplots. Here we can see that these two samples are outliers but otherwise both conditions show reasonably similar distributions for this parameter.

Users can also explore potential sources of contamination from reads mapped to viral and bacterial genomes. Figure 15D shows that all samples suffer from rather high percentages of contamination reads. All samples have more bacterial/viral reads than 75% of all animal samples in the reference set. This range of values indicates serious contamination with the possible exception of cervical cancer samples, where this might be caused by sample extraction or even have a role in the disease [288].

Finally, Figure 15E shows the library complexity of different tissues from *Takifugu rubripes* [289]. While the top expressed microRNA in intestine picks up 47.1% of all reads (percentile 88.4), in ovary this figure drops to 8.3% which corresponds to percentile 2.6. In ovary, eight microRNA sum

up over 50% of all miRNA expression while in intestine it takes only two to reach the same percentage which indicates a higher complexity of miRNA expression in germ cells. Furthermore, ovary and testis exhibit much lower percentages of miRNAs. This might be related to their larger repertoire of small RNAs in germ cells (21) which automatically would lead to a lower relative fraction of microRNAs in those samples.

6.6 CONCLUSION

We present a user-friendly web server for the comparative quality control of miRNA-seq data that can be useful in several scenarios: to identify low-quality samples that should be excluded from downstream analysis; to reveal systematic errors in order to improve the library preparation process, something especially relevant for pilot studies; and finally, to provide external quality validation for datasets so it can be used as a standard proof of quality.

mirnaQC provides several output tables and visualisations for a total of 34 quality attributes which allow users to rank their results against a large corpus of comparable samples. In this way, no absolute thresholds need to be applied and the user can evaluate their sequencing data based on percentiles. Future developments include new types of analysis and improved visualisations intended to detect confounding variables related to quality issues that can affect downstream steps. Additionally, a dockerized version of the tool will be made available so the pipeline can be run locally or in computing clusters.

6.7 DATA AVAILABILITY

https://arn.ugr.es/mirnaqc/

6.8 ACKNOWLEDGEMENTS

The authors acknowledge the usage of the computational infrastructure of the Computational Epigenomics Lab of the University of Granada.

6.9 FUNDING

This work was supported by European Union [765492]; Spanish Government [AGL2017-88702-C2-2-R] to M.H.; Consejería de Economía, Conocimiento, Empresas y Universidad de la Junta de Andalucía and European Regional Development Funds (ERDF) [SOMM17-6109, UCE-PP2017-3] to J.A.M. and M.H.; Instituto de Salud Carlos III, ERDF funds [PIE16/00045] to J.A.M.; Chair 'Doctors Galera-Requena in cancer stem cell research' (to J.A.M.); Instituto de Salud Carlos III [IFI16/00041] to E.A. Funding for open access charges: Excellence Research Unit "Modelling Nature" (MNat) [SOMM17-6109].

Chapter 7

An exploration of the circulating miRNAs in melanoma patients to propose candidate biomarkers for early tumor detection

7.1 INTRODUCTION

Melanoma is a type of cancer that originates from melanocytes, the cells that produce pigmentation. This type of cancer most typically occurs in the skin but can also develop in other parts of the body such as the eyes [290]. Globally, melanoma is only the 19th most frequent type of cancer but its incidence is much higher in Australia and New Zealand [291], where people of fair skin are exposed to UV (Ultraviolet) light, or in Northern Europe where the population, also of fair skin, engages in winter or sun-seeking holidays [292]. Since melanoma is a public health concern in those countries, the increased diagnosis statistics may also be a result of improved detection [292]. In fact, melanoma was the third most common cancer diagnosis in Australia in 2020, only behind breast and prostate cancer, and the most common among people between the ages 20-39 [293]. The risk is also increased among the elderly, as there may be a lag of up to 40 years between the exposure to UV and the onset of the disease [292].

As it happens with virtually every type of cancer, early detection and treatment of melanoma improves its prognosis. In line with this, survival rates in melanoma display staggering differences between stages: 5-year survival rate in the US is 99% for localized disease (Stages 0, I and II), 66% in the case of regional spread and only 27% for patients with distant spread [294]. Similar trends have been determined for 10-year survival. It should be noted that survival rates at later stages have dramatically increased over the past few years (they used to be 10-20%) thanks to newly available therapies including immunotherapy [295] and targeted therapies [296] that are most effective when applied early.

In this context, it becomes apparent that early diagnosis tools can be of great help to improve melanoma management. Current melanoma diagnosis is based on detection of suspicious moles or other skin lesions that are resected and assessed by pathology specialists. This may be impractical in people with many moles or people of darker skin where lesions are harder to detect. Some tissue biomarkers are of clinical interest and can be used to assess the prognosis but there are no established circulating biomarkers since studied candidates are either of little specificity such as LDH (Lactate dehydrogenase) or only useful to assess prognosis at late stages [297].

Liquid biopsy is among the most practical strategies [298] to introduce and comply with screening programs and would dramatically decrease the lag between detection and diagnosis in potential melanoma patients, which currently need specialized sample processing and assessment. Previous studies have identified potential circulating miRNA biomarker panels in melanoma patients, but these efforts mostly relied on technologies like mi-

croarray [299] or qPCR [167] which cannot be reanalyzed using new miRNA references or additional RNA libraries and that cannot be used to assess changes in isomiR frequencies or proportions.

In this work, we explored the miRNA profiles of 26 melanoma patients of all stages by means of miRNA-seq. Several potential miRNA-based biomarkers, which partially overlap with results from previous similar studies, were identified as good predictors of tumor presence or early stage. We also explored the impact of uridylation to reveal some miRNAs with NTA-U levels that correlate with disease progression. Further work should include the validation of these biomarkers in a larger independent cohort.

7.2 MATERIAL AND METHODS

7.2.1 Sample collection

Blood samples were obtained from the Oncology Service at the University Hospital Virgen de las Nieves, Granada and University Hospital San Cecilio, Granada (Spain). The study was approved by the ethics committee of both hospitals (register number: 32140085) and all clinical research was conducted according to the principles reflected in the Declaration of Helsinki ('Ethical Principles for Medical Research Involving Human Subjects'). Written consent was obtained from all controls and patients prior to their enrolment in the study [300]. In total, 31 samples were collected from healthy subjects and melanoma stages 0, I, II, III and IV (See Table 2). Blood samples were collected in BD vacutainer SSTII advanced tubes (Becton Dickinson, Franklin Lakes, NJ, USA), incubated at room temper-

ature for 30 min and centrifuged for 10 min at 1400g. The resulting serum was stored at −80°C until submission to sequencing services.

Table 2: Summary of clinical details of patient recruited in the study.

Stage	n	Age (median)	Sex (M/F)
Control	5	Unknown	Unknown
Melanoma - 0	5	74	3/2
Melanoma - I	5	48	4/1
Melanoma - II	5	68	3/2
Melanoma - III	4	50	3/1
Melanoma - IV	7	63	4/3

7.2.2 Small RNA sequencing

Serum samples were submitted for sequencing to the Genomics Unit of the Germans Trias i Pujol Research Institute (IGTP). The samples were prepped with the TruSeq small RNA Illumina protocol (2014 version) adapted to serum samples with automated pooled library size selection using Pippin prep. DNA libraries were generated and indexed prior to size selection. Quality control was performed after size selection using Bioanalyzer (Agilent) with DNA 1000 assay chip. Single read 50nt libraries were sequenced on a cBOT-HiSeq-2500. The number of sequenced reads per sample ranged from 7M to 59M (See Figure 16).

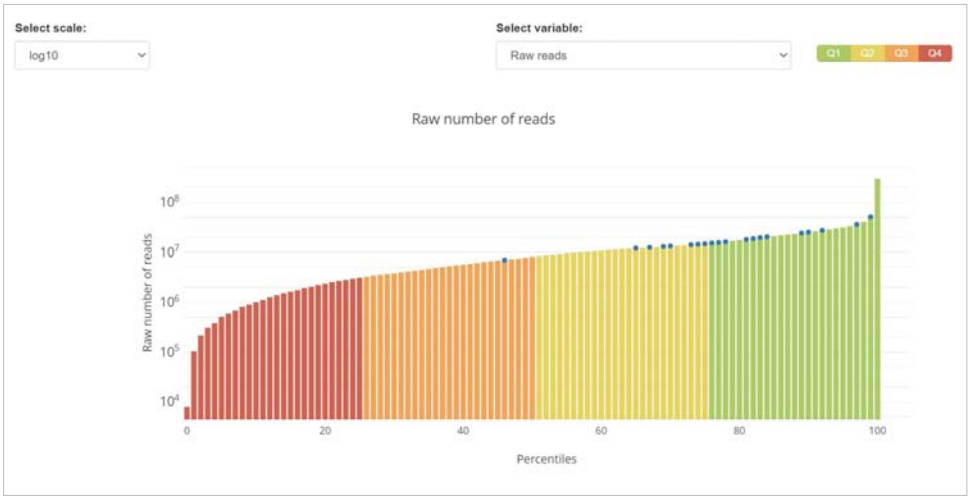


Figure 16: Number of reads sequenced per experiment (dots) in the context of all liquid biopsy samples in SRA (bars). The outlier in percentie 37 is MIII-1 (generated using *mirnaQC*).

7.2.3 Quality control of FASTQ files

Prior to preprocessing and analysis of the reads, quality control of the fastq files was carried out using *mirnaQC* [231]. Briefly, fastq files were uploaded and processed using the mirnaQC webserver and the resulting quality metrics were ranked using all liquid biopsy samples in the database as background. Outliers were detected by checking basic quality features presented in the overview of the tool and by further exploration of the problematic features with PCA of expression values (50 most expressed miRNAs) combined with color labels derived from the rank (see Figure17). Outliers were removed from the downstream differential expression analysis.

7.2.4 Pre-processing and mapping

Fastq files were preprocessed with *sRNAbench* [218] using parameter protocol=Illumina. Reads were aligned on library mode using the following references hierarchically mapped in this order: miRNAs (miRBase release 22.1), tRNA (GtRNAdb 2.0), Y RNA (from RNA central), remaining ncRNAs from Ensembl, remaining ncRNAs from RNAcentral, mRNA fragments (cDNA from Ensembl), vertebrates viral genomes (retrieved from NCBI), bacterial genome collection (retrieved from NCBI). isomiRs were also profiled and classified using *sRNAbench*'s hierarchy including adenylations and uridylations (NTA-A and NTA-U).

7.2.5 Differential expression

Differential expression analysis was performed using *DESeq* [237], *DESeq2* [238], *edgeR* [235], *NOISeq* and Student's t-test [234] on RPM (Reads Per Million). Different comparisons were performed with sRNAde and then summarized using a python script. To correct for multiple testing, Benjamini–Hochberg procedure was applied so adjusted p-value and FDR (False Discovery Rate) are used interchangeably. Outlier samples were removed from the analysis to avoid statistical power reduction.

7.2.6 Differential uridylation

Uridylation levels per miRNA were calculated using *sRNAbench* with parameter isoMiR=true. A one-way ANOVA test (ANalysis Of VAriance) was performed on each miRNA to detect if differences in average group

urydilation were significant. Samples were grouped according to clinical cancer stage (Control, 0, I, II, III, IV).

7.3 RESULTS

7.3.1 Quality Control of miRNA-seq libraries

Overview of the most relevant quality features using the heatmap tool revealed 7 samples with high amounts of short fragments (Q4 compared to human liquid samples in SRA). One of them was a clear outlier in terms of quality (MIII-3) that displayed lower values in two extra features: Percentage of reads in miRNA peak (21-23nt) and a higher standard deviation of the length distribution of miRNA reads. Upon further inspection in the "Read length distribution" section of the tool we could confirm that the skewness of the distribution was also lower (flatter) for this sample. Combined, these signs indicate possible RNA degradation. Exploration of the PCA from miRNA expression values revealed 3 further outliers: MIII-2, MIV-7 and MIV-5. These samples displayed an increased amount of ultrashort fragments (discarded previous to the analysis) which may have affected their expression pattern (see Figure 17). This artifact could also be an indication of RNA degradation or, if present across all cDNA libraries processed together, of imperfect size selection.

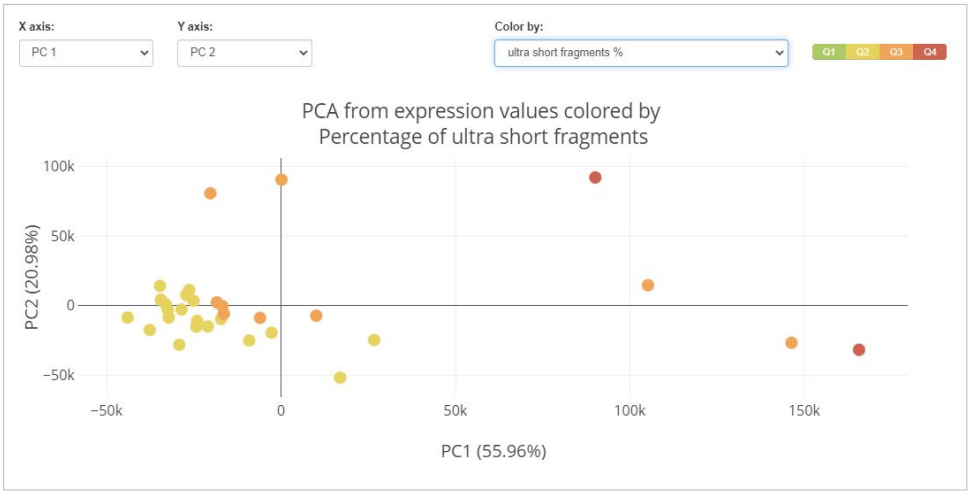


Figure 17: PCA of the top50 most expressed miRNAs. Samples are colored using the *mirnaQC* quartile for "percentage of ultrashort reads". Four outliers can be observed.

7.3.2 RNA categories distribution

On average, mature miRNAs accounted for the largest fraction of small RNAs present in each sample (∼33%), followed by Y RNA (∼23%). tRNA fragments and bacterial sequences also accounted for a significant amount of reads ∼14% and ∼5% respectively. The remaining categories were less than 5% of reads each (See Figure 18). Some samples had a remarkably smaller miRNA fraction than their counterparts: MI-1, MIII-2, MIII-3 and MIV-5, three of which were already spotted as outliers in the quality control step.

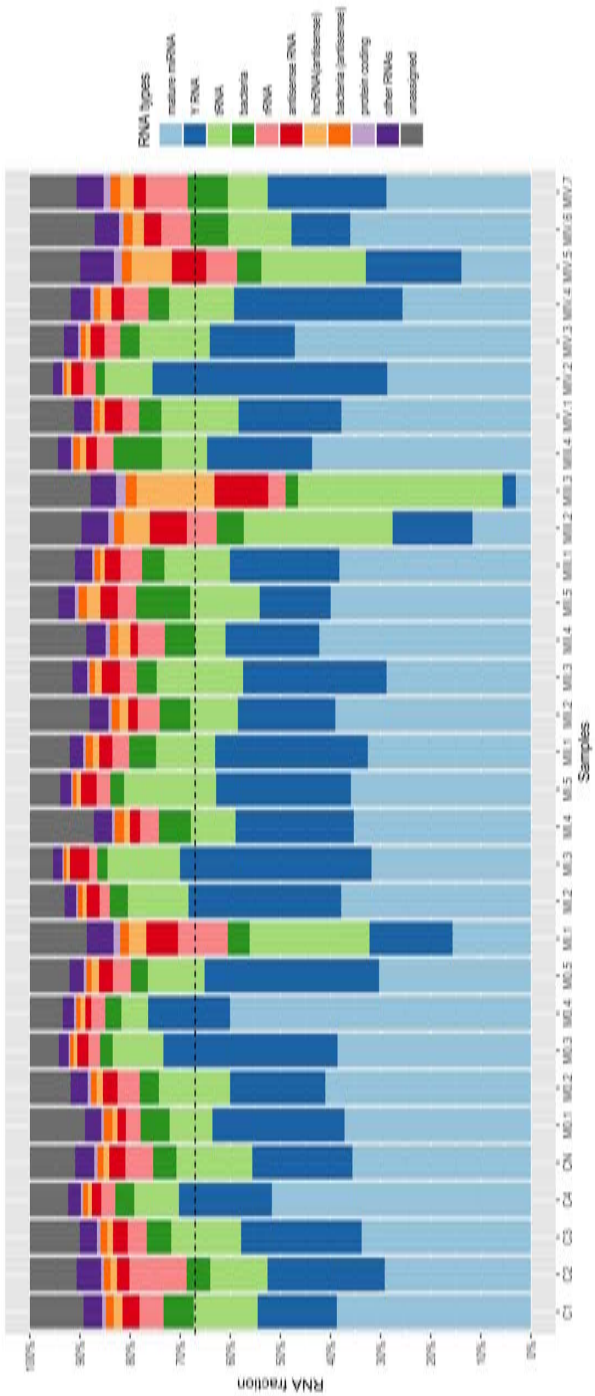


Figure 18: Fraction of reads corresponding to each RNA category

7.3.3 miRNA expression

On average, 33% of mapping reads were assigned to different miRNAs (range 3-60%). The number of mature miRNA references receiving mappings ranged from 421 to 1282 (average 1116). This did not seem to be related to sequencing depth (see Figure 19A) and again, samples outside the expected ranges overlapped with the outliers described before. A similar trend can be observed when analyzing miRNA-mapping reads (see Figure 19B).

The most abundant miRNA took on average 25% of the mapping reads, the top 5 took 49%, the top 10 63%, the top20 78% and the top 50 93%. Some of the outliers show very different behavior in the most expressed miRNAs (see Figure 20). For instance MI-1 only has 9% of reads mapping to miR-486-5p and MIII-2, MIV-5 and MIV-7 have abnormally low miR-92a-3p values.

All samples shared only two miRNAs among their 5 most expressed: miR-486-50 and miR-22-3p (same for the top 10 most expressed). The number of shared miRNAs rises to 5 among the top 20 and to 31 among the top 50. Finally, all samples have 64 miRNAs in common among the top 100. This same analysis was replicated after removing the outliers detected before, which increased the percentage of shared miRNAs among samples (see Table 3).

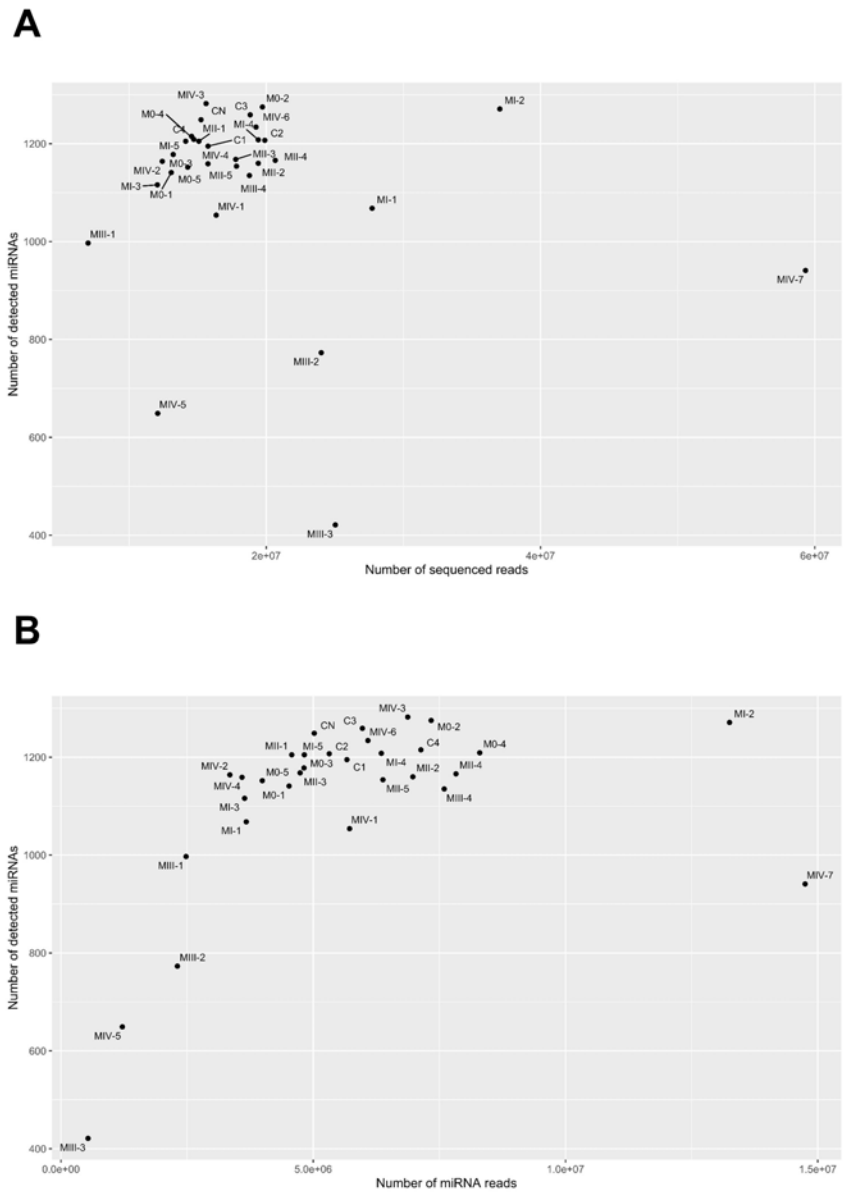


Figure 19: Number of mature miRNAs detected per sample compared to total number of sequenced reads (A) and total number of miRNA mapping reads.

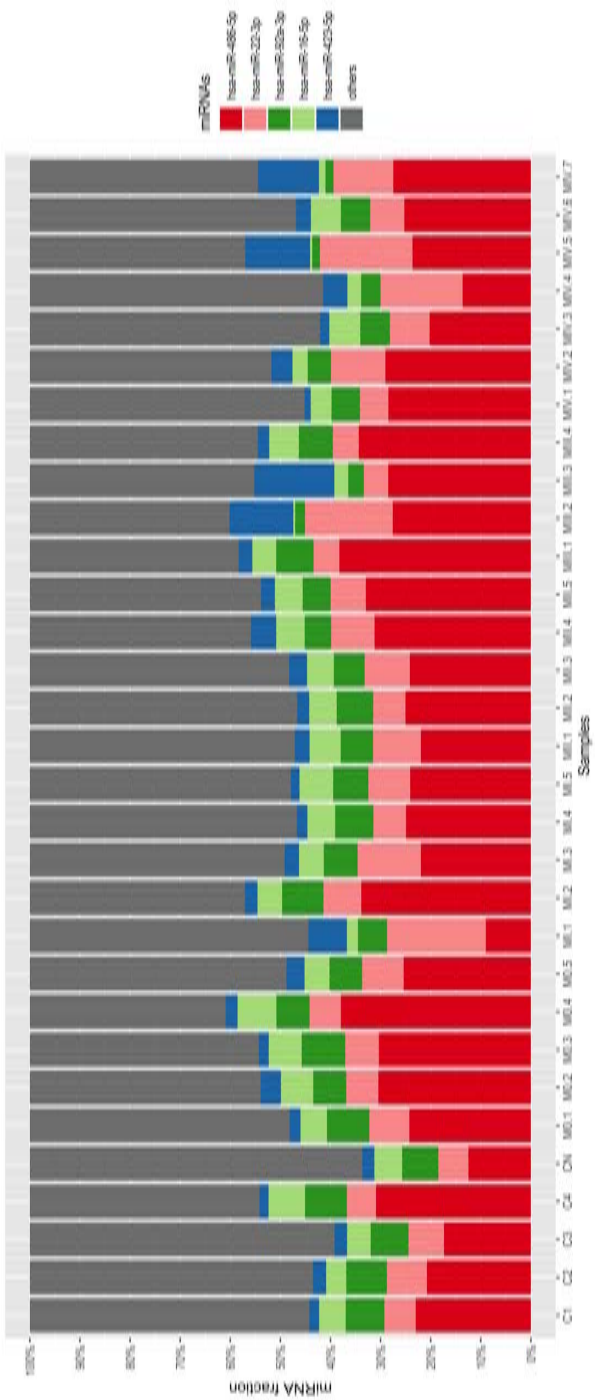


Figure 20: Fraction of reads corresponding to the 5 most expressed miRNAs

Table 3: Number of most expressed miRNAs (N) shared by all samples in the study before and after removing the outliers.

N	Shared all samples	All samples (%)	No Outliers	No Outliers (%)
5	2	40%	4	80%
10	2	20%	4	40%
20	5	25%	8	40%
50	31	62%	38	76%
100	64	64%	78	78%

7.3.4 Differential expression

Controls versus cancer (melanoma stages 0, I, II, III, IV)

A total of 143 different miRNAs were differentially expressed between both conditions according to at least one DE method (p-val < 0.05) and 21 were differentially expressed according to 4 out of 5 methods, some are shown in Figure 21). If we consider FDR < 0.05, only 4 miRNAs meet the requirement: hsa-miR-760, hsa-miR-658, hsa-miR-4508 (underexpressed, called by *DESeq2*) and hsa-miR-664a-3p (overexpressed, called by t-test).

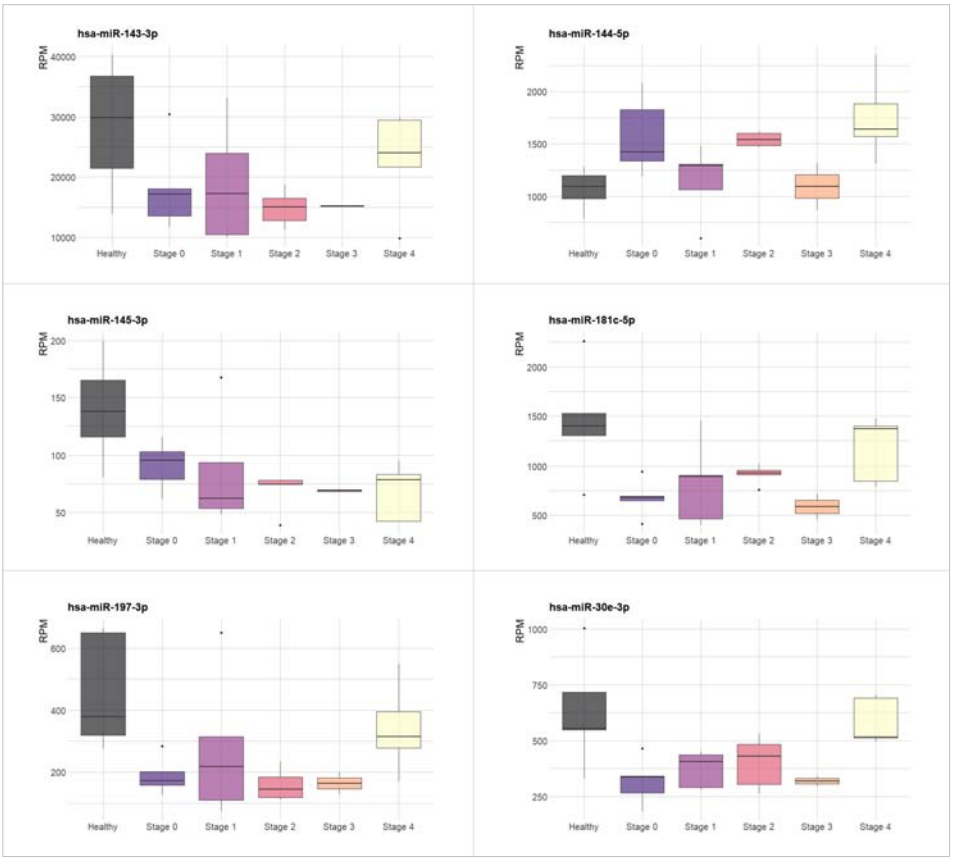


Figure 21: Differentially expressed miRNAs (healthy control vs all cancer stages) called by 4 out of the 5 DE methods. 21 miRNAs were differentially expressed ($p < 0.05$) but only 6 were selected based on relative abundance (higher is favored) and inter-group variability (more homogeneous cancer groups were favored)

Control vs stages 0, I, II (Early detection)

164 miRNAs were differentially expressed between both groups, 41 of which were predicted by 4 methods and 35 by 3 methods. After FDR correction, 22 miRNAs were still significantly DE by at least one method. Among these, hsa-miR-664a-3p was detected by 3 different methods (also a marker in the previous analysis) as well as has-miR-199b-5p, predicted by 2 methods and a candidate marker to discriminate between healthy and cancer. Selected differentially expressed miRNAs are shown in Figure 22.

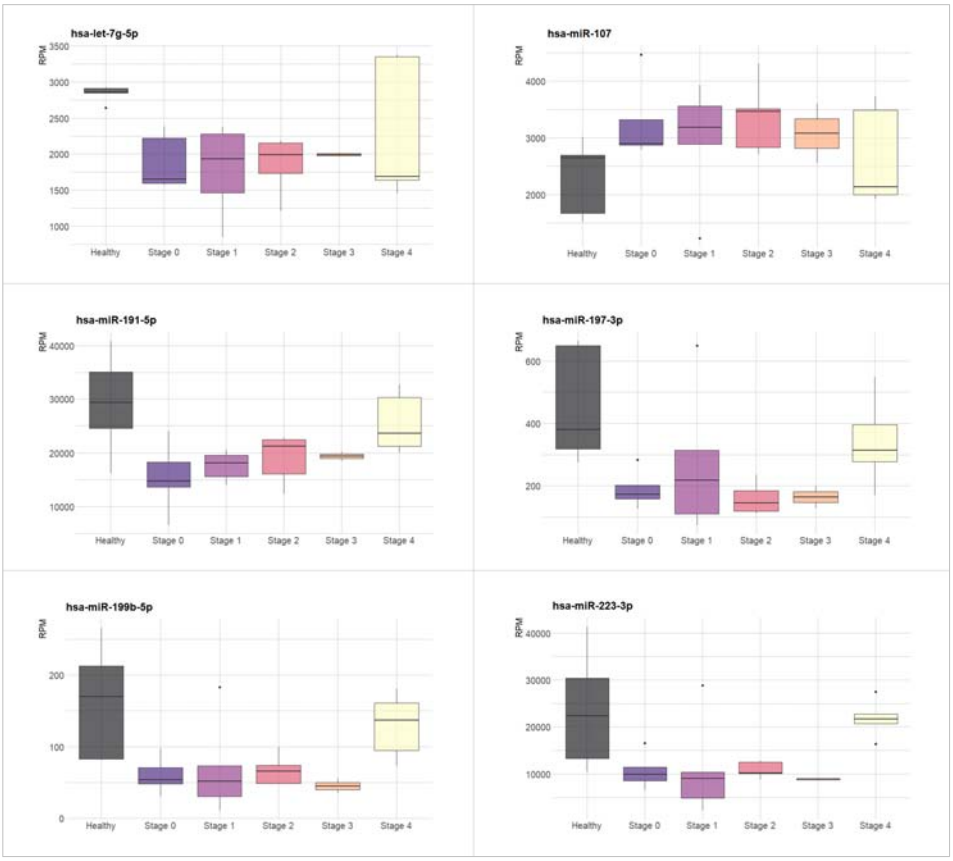


Figure 22: Differentially expressed miRNAs (healthy control vs stages 0, 1 and 2) called by 4 out of the 5 DE methods. 41 miRNAs were differentially expressed ($p < 0.05$) but only 6 were selected based on relative abundance (higher is favored) and inter-group variability (more homogeneous cancer groups were favored

7.3.5 isomiR analysis

For each sample, the fraction of each isomiR category was calculated to study if the stage had any impact on miRNA processing (See Figure 23). No effects can be appreciated that relate to progression of melanoma.

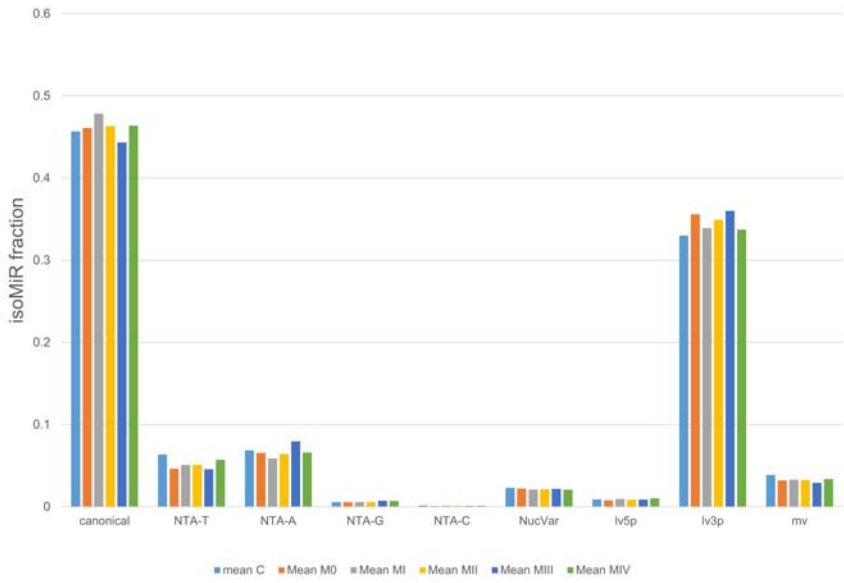


Figure 23: Fraction of different isomiR types. Samples are grouped according to clinical cancer stage

Differential uridylation

Differences in uridylation at the level of single miRNAs were also explored as a biomarker source. Those miRNAs that displayed some tendency correlating to progression are shown in Figure 24. Among those, miR-127-3p stood out (p-value < 0.002743016, ANOVA), a miRNA that has been de-

scribed to play a role in melanoma progression [301, 302] and other types of cancer [303, 304]. miR-370 has actually been reported to increase its expression with melanoma progression [305], effectively working as an oncogene. miR-30e-3p is also of particular interest since it's also differentially expressed (3/5 methods) between control and cancer patients, although it shows a very small difference (from slightly over 10% uridylation in earlier stages to slightly under 10% in later stages).

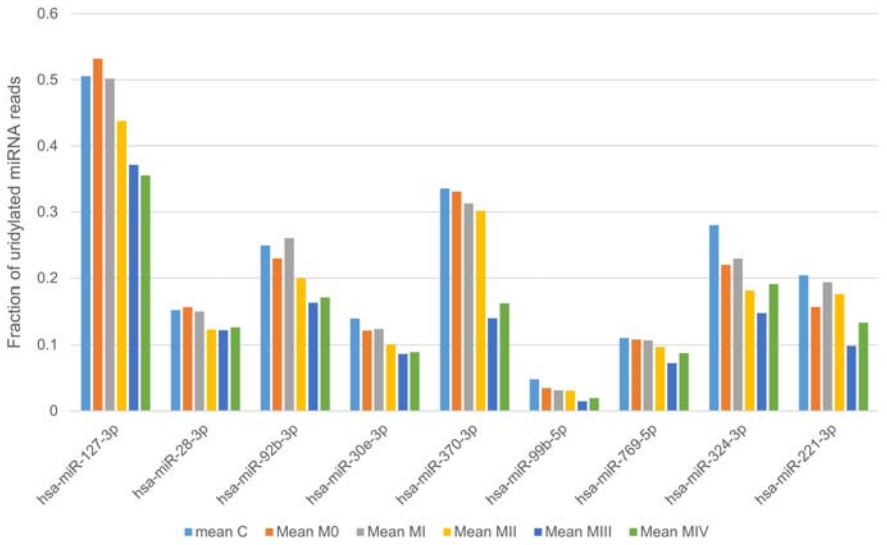


Figure 24: Fraction of uridylated (NTA-U) reads for selected miRNAs. Samples are grouped according to clinical cancer stage.

7.4 DISCUSSION

Liquid biopsy has emerged as a promising approach for early cancer detection. Types of cancer with a wide survival gap between early and late stages such as melanoma, where stage I diagnoses are 4 times more likely to survive

than stage IV patients, will benefit the most from these techniques since they provide an opportunity of early treatment where chances of patient survival are maximal. Additionally, the time delay between sampling and diagnosis would dramatically decrease as a consequence of applying liquid biopsy methods which merely require a simple blood extraction. Among the circulating biological materials, cell-free miRNAs have gathered the most attention [171] probably because of their relative stability in bodily fluids.

In this study we explored the abundance of circulating miRNAs currently annotated in miRBase in a dataset of 26 melanoma patients using miRNA-seq. Melanoma patients were evenly distributed across all clinical stages (0, I, II, III, IV) and 5 control samples were obtained from healthy volunteers. After performing two differential expression comparisons, control vs cancer and control vs early stages, we identified 143 and 164 miRNAs, respectively, that were significantly over-/underexpressed according to at least 4 out of 5 DE methods. We took this consensus-based approach to identify DE miRNAs because we found that multiple testing correction was too severe and close to zero miRNAs were detected as differentially expressed after applying FDR. This way, however, we were able to increase confidence on the analysis without losing much statistical power. An enrichment analysis (Gene Set Enrichment Analysis) was performed using miEAA2 [306] and the DE miRNAs as input. This analysis revealed that many cancer terms were significantly enriched including medulloblastoma, bladder cancer, NSCLC, rectum adenocarcinoma, lymphoma and, most importantly, melanoma (it was the disease with the 11th lowest p-value). It is also noteworthy that some of the miRNAs identified as DE had been previously reported by similar studies. For instance, hsa-miR-664b-5p and

hsa-miR-145-5p were also described as deregulated by a previous study using microarrays [299]. Interestingly, loss or mutation of hsa-miR-664b-5p has been linked to proliferation in the context of melanoma [307] and other malignancies [308]. Nevertheless, conclusions drawn from this set of molecules should be extremely cautious since the relative abundance of each of these miRNAs was rather low: they were all below 50,000 RPM (equivalent to 0.5%) and frequently below 10,000 or even 1,000 RPM. In any case, these biomarkers or their combination into a melanoma diagnosis panel should be confirmed in a validation study using an independent cohort.

We also explored the distribution of the isomiR repertoire across the patients. Particularly, we focused on NTA-U (non-templated uridylation), a type of post-transcriptional modification that has been linked to arm switching [309], miRNA degradation [310] and alternative targeting [230]. miR-127-3p was the only miRNA with significant changes in the mean fraction of uridylated reads per group (p-value < 0.0027, ANOVA) although its levels were not differentially expressed in the comparisons explored. miR-127-3p has been reported to behave as a tumor suppressor in several types of cancer including epithelial ovarian cancer [311] , glioblastoma [312], gastric cancer [313] and osteosarcoma [303]. More importantly, miR-127-3p is underexpressed in melanoma cells [302] and serum of patients [314] and its overexpression was reported to repress melanoma progression in in vivo [302] and in vitro [301] experiments. Although the role of miR-127-3p in the progression of melanoma is yet to be fully understood, it targets the Protein delta homolog 1 (DLK1) which has been described to modulate NOTCH1-dependent proliferation in melanoma cells [315] and miR-127-3p

restoration reduces cell proliferation in melanoma models [302]. According to our analysis, miR-127-3p uridylation decreases from around 50% in controls and early stages of melanoma (0, I) to around 35% for later stages (III and IV). This could hint an increase in the turnover of this mature miRNA but it would be premature to link it to any physiological function since cancer tissue miRNA profiles were not available. Finally, it is worth mentioning that miR-127-3p has been used in expression panels for the diagnosis of melanoma [316] and other types of cancer such as lymphoma [317].

Although the biomarkers proposed here seem promising, it must be admitted that this study is limited by the size of the cohort (n=31). Assessment of any kind of classification produced using this small dataset will inevitably lead to an overestimation of its accuracy, a common problem known as overfitting. To avoid such a pitfall, an independent validation cohort is necessary. Ideally, this approach would allow to compose a panel of biomarkers using this dataset that can correctly classify patients in the validation cohort. Furthermore, we observed that very few miRNAs took a large fraction of the miRNA reads in all samples which could be caused by different artifacts and can lead to erroneous quantification. Validation using a different technique that is prone to different bias, such as RT-PCR, would consequently increase the confidence in the biomarkers proposed.

In summary, we have proposed several miRNA-based biomarkers for detection of melanoma using NGS data from 26 patients. Many of the proposed markers had been described to have a role either in the progression of melanoma or of some other type of cancer. Furthermore, some of them have even been proposed as biomarkers in several cancer diagnosis panels that

are applied to either blood or tissue samples. Even though these biomarkers seem promising, validation in an independent cohort is still necessary. This work provides a starting point for the development of miRNA-based liquid biopsy panels for early detection of melanoma.

Chapter 8

Conclusions

1. As part of this thesis, different bioinformatics methods were developed and implemented into user-friendly tools in order to make analysis of small RNA NGS data more reliable and reproducible. Special focus was set on the quality control and analysis of samples from bodily fluids which can be highly problematic, especially if their input is low. These tools are: 1) *sRNAbench*, an updated pipeline and webserver for the analysis of miRNA-seq profiles; 2) *liqDB*, a manually curated database of liquid biopsy miRNA-seq profiles and 3) *mirnaQC*, a webserver for comparative quality control of miRNA-seq data.

2. Some of the implemented methods required a large number of miRNA-seq experiments and, in order to obtain and process suitable samples, some sort of workflow was needed. We developed a pipeline that can retrieve updated metadata from SRA to select likely miRNA-seq samples, including potentially misclassified experiments that mention miRNA in their title or description. Once said samples have been downloaded, the relevant information is derived from the metadata or inferred from the sample reads.

3. Accurate inference of library preparation protocol can be achieved by means of a multi-step iterative method. This information is crucial for the appropriate preprocessing of miRNA-seq samples but often missing or poorly annotated in SRA metadata. A tool was implemented to detect any commercially available small RNA protocol in addition

to custom protocols containing random adapters and/or UMIs. The algorithm is based on maximization of miRNA detection in a subset of reads.

4. Quality control of miRNA-seq experiments is a nontrivial problem. We analyzed over 30 quality parameters across more than 30,000 publicly available miRNA-seq samples to determine how their distribution affected the resulting miRNA profiles. We stored the distribution of these parameters into a database and organized the quality features into categories: sequencing yield, library complexity, library quality, putative contamination, length distribution and sequencing quality. The selected parameters are still to be jointly considered and they only provide limited explanations for unwanted technical variability. Documentation is provided to help users identify specific issues with their samples.

5. Visualizations and quality features provided in *mirnaQC* enable identification of outliers and point to likely technical causes behind the issues. Rather than arbitrarily decided thresholds, these features are provided as percentiles so users can interpret them in the context of a large corpus of uniformly processed samples. In the chapter that describes *mirnaQC*, several examples of successful application of this process are discussed.

6. Quality control of miRNA-seq data is context-dependent. We were first struck by this fact when trying to identify outliers in our in-house liquid biopsy data: when compared to the whole corpus of experiments all of our samples seemed of very poor quality according

to most parameters, which therefore appeared to be of little to no use. This was remedied by implementing subset comparisons, which allows to compare only to samples selected by the user. Subsets include samples from the same species, samples with the same library preparation protocol and samples from bodily fluids.

7. Small RNA sequencing data from liquid biopsies can be affected by multiple confounding variables including sequencing artifacts such as ligation bias. Identification of such factors can only be achieved by systematic comparison of large amounts of sequencing data covering all those variables. In *liqDB* we compiled, processed and manually curated the 31 liquid biopsy miRNA-seq studies publicly available at the time into a freely accessible database for data download and reanalysis.

8. Both *liqDB* and *mirnaQC* required data processing by *sRNAbench*. To accommodate new protocols and integrate some improvements, *sRNAbench* was reimplemented and dockerized, which allows easy and quick portability between computers. New developments include one-step launching of multiple samples, updated small RNA references and protocols, the redesign of several result pages including new visualizations and the possibility to process spike-in sequences.

9. Differential expression of miRNA transcripts is an understudied task and most researchers rely on differential expression methods modeled on mRNA sequencing data. Therefore, the differential expression module of sRNAtoolbox was improved to incorporate the 5 most frequently used methods together with a consensus set, a strategy that

allows prioritization of candidate biomarkers. Additionally, the detection of significant differences in processing patterns among groups was also implemented to take advantage of the extra layer of information found in isomiRs and isomiR fractions.

10. To show the usefulness of this approach, we applied the methods developed in this thesis to a set of liquid biopsy samples from melanoma patients. Several miRNAs were found to be differentially abundant between healthy controls and melanoma patients according to different DE methods. Furthermore, some miRNAs also exhibited potential as early melanoma biomarkers.

11. Finally, a few biomarkers were proposed based on their uridylation pattern. Among these, miR-127-3p showed remarkable potential because of its relative abundance (3000 RPM), a consistently decreasing pattern (from 50% uridylation in controls to 35% in stage IV) and being statistically significant. Besides, miR-127-3p was also previously linked to melanoma and its overexpression has been reported to repress cancer progression.

Publications

Publications of this thesis

- *Ernesto Aparicio-Puerta, Ricardo Lebrón, Antonio Rueda, Cristina Gómez-Martín, Stavros Giannoukakos, David Jaspez, José María Medina, Andreja Zubkovic, Igor Jurak, Bastian Fromm, Juan Antonio Marchal, José Oliver, Michael Hackenberg* **sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression** Nucleic Acids Research, 2019, doi: 10.1093/nar/gkz415.

- *Ernesto Aparicio-Puerta, David Jáspez, Ricardo Lebrón, Danijela Koppers-Lalic, Juan A Marchal, Michael Hackenberg* **liqDB: a small-RNAseq knowledge discovery database for liquid biopsy studies** Nucleic Acids Research, 2019, doi: 10.1093/nar/gky981.

- *Ernesto Aparicio-Puerta, Cristina Gómez-Martín, Stavros Giannoukakos, José María Medina, Juan Antonio Marchal, Michael Hackenberg* **mirnaQC: a webserver for comparative quality control of miRNA-seq data** Nucleic Acids Research, 2020, doi: 10.1093/nar/gkaa452.

Other publications

- *Thomas Desvignes, Phillipe Loher, Karen Eilbeck, Jeffery Ma, Gianvito Urgese, Bastian Fromm, Jason Sydes, Ernesto Aparicio-Puerta,*

Victor Barrera, Roderic Espín, Florian Thibord, Xavier Bofill-De Ros, Eric Londin, Aristeidis G Telonis, Elisa Ficarra, Marc R Friedländer, John H Postlethwait, Isidore Rigoutsos, Michael Hackenberg, Ioannis S Vlachos, Marc K Halushka, Lorena Pantano **Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API** Bioinformatics, 2019, doi: 10.1093/bioinformatics/btz675.

- *Bastian Fromm, Diana Domanska, Eirik Høye, Vladimir Ovchinnikov, Wenjing Kang, Ernesto Aparicio-Puerta, Morten Johansen, Kjersti Flatmark, Anthony Mathelier, Eivind Hovig, Michael Hackenberg, Marc R Friedländer, Kevin J Peterson* **MirGeneDB 2.0: the metazoan microRNA complement** Nucleic Acids Research, 2019, doi: 10.1093/nar/gkz885.

- *Chantal Scheepbouwer, Kayla Borland, Ernesto Aparicio-Puerta, Heleen Verschueren, Laurine Wedekind, Jip Ramaker, Branko Misovic, Mathilde CM Kouwenhoven, David Noske, Peter Vandertop, Pieter Wesseling, Tom Wurdinger, Michael Hackenberg, Stefanie Kellner, Danijela Koppers-Lalic* **The epitranscriptomic code in LGG: Metabolically reprogrammed IDH-mutant gliomas alter tRNA modification landscape** Neuro-Oncology, 2019, doi: 10.1093/neuonc/noz175.462.

- *Ernesto Aparicio-Puerta, Bastian Fromm, Michael Hackenberg, Marc K. Halushka* **In Silico Analysis of Micro-RNA Sequencing Data. Chapter 13 of RNA Bioinformatics** Methods in Molecular Biology, vol 2284. Humana, New York, NY, 2021, doi: 10.1007/978-

1-0716-1307-8_13.

- *Fabian Kern, Lena Krammes, Karin Danz, Caroline Diener, Tim Kehl, Oliver Küchler, Tobias Fehlmann, Mustafa Kahraman, Stefanie Rheinheimer, Ernesto Aparicio-Puerta, Sylvia Wagner, Nicole Ludwig, Christina Backes, Hans-Peter Lenhof, Hagen von Briesen, Martin Hart, Andreas Keller, Eckart Meese* **Validation of human microRNA target pathways enables evaluation of target prediction tools** Nucleic Acids Research, 2021, doi: 10.1093/nar/gkaa1161.

- *Fabian Kern∗ , Ernesto Aparicio-Puerta∗ , Yongping Li∗ , Tobias Fehlmann, Tim Kehl, Viktoria Wagner, Kamalika Ray, Nicole Ludwig, Hans-Peter Lenhof, Eckart Meese, Andreas Keller* **miRTargetLink 2.0—interactive miRNA target gene and target pathway networks** Nucleic Acids Research, 2021, doi: 110.1093/nar/gkab297.

  ∗ The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

- *Cristina Gómez-Martín, Ernesto Aparicio-Puerta, José M. Medina, Guillermo Barturen, José L. Oliver, Michael Hackenberg.* **geno5mC: A Database to Explore the Association between Genetic Variation (SNPs) and CpG Methylation in the Human Genome** Journal of Molecular Biology, 2020, 166709, doi: 10.1016/j.jmb.2020.11.008.

- *Esther E. E. Drees, Margaretha G. M. Roemer, Nils J. Groenewegen, Jennifer Perez-Boza, Monique A. J. van Eijndhoven, Leah I.*

Prins, Sandra A. W. M. Verkuijlen, Xuan-Mai Tran, Julia Driessen, G. J. C. Zwezerijnen, Phylicia Stathi, Kevin Mol, Joey J. J. P. Karregat, Aikaterini Kalantidou, Andrea Vallés-Martí, T.J. Molenaar, Ernesto Aparicio-Puerta, Erik van Dijk, Bauke Ylstra, Catharina G. M. Groothuis-Oudshoorn, Michael Hackenberg, Daphne de Jong, Josée M. Zijlstra, D. Michiel Pegtel **Extracellular vesicle miRNA predict FDG-PET status in patients with classical Hodgkin Lymphoma** Journal of Extracellular Vesicles, 2021, 166709, doi: 10.1002/jev2.12121.

List of Figures

List of Tables

Acknowledgments

Dicen que es de bien nacidos ser agradecidos. No sé si se podrá decir que fui "bien nacido" (teniendo en cuenta que soy sietemesino y que me tuvieron que madurar un poquito en la incubadora) pero desde luego voy a intentar ser todo lo justo posible con la gente gracias a la cual puedo ahora escribir estas palabras (siempre dentro de lo limitado de mi cerebro prematuro). Al que la introducción le haya parecido un rollazo que se agarre a la silla.

El protocolo dicta que debo empezar por mis directores, algo que haré encantado y que considero más que justo. He de decir que una de las cosas que me sorprendieron bastante al comenzar la tesis fue la gran cantidad de colaboraciones que mantienen estos dos culos inquietos. A mí me parecía insostenible, no solo por la cantidad de trabajo que les suponía a ellos, sino porque, si yo fuera colaborador suyo, ¿estaría dispuesto a lidiar con tan poca disponibilidad por su parte? ¿Aceptaría que mi proyecto fuera "uno más" de tantos para ellos? Con el paso de los meses me fui dando cuenta de que mantienen tan apretada lista de colaboraciones a base de mucha dedicación y conocimiento pero también de algo casi tan importante: actitud. Mi sensación desde el principio fue de "buen rollo" con ellos, algo que sin duda favorece la creación y sostenimiento de vínculos con otros investigadores. Recuerdo perfectamente entrar a reuniones con cualquiera de ellos malhumorado o con dudas y salir siempre reconfortado y con la sensación de que tenía mucha suerte de poder trabajar con ellos (nótese que no "para" ellos). A cualquiera que por casualidad esté leyendo estas líneas y planteándose si trabajar para o con cualquiera de ellos: ¡no lo dudes!

Juan, muchas gracias por darme la oportunidad sin conocerme y por haberme apoyado SIEMPRE en cada cosa que he intentado. De no ser por tu aliento y optimismo seguramente no me habría presentado a muchas convocatorias que parecían inaccesibles. Me he sentido acogido y valorado en tu grupo a pesar de que fuera el tío raro de los ordenadores. Aunque pronto me vaya para continuar mi carrera en otra parte, espero que cuentes conmigo si en algún momento crees que te puedo ser útil. Gracias de nuevo, ha sido un placer trabajar contigo.

Michael ¿qué decirte? Cuando empezaba la tesis alguien me habló de la relación quasi paterno-filial que se desarrollaba entre director y doctorando si todo iba bien. En mi caso siempre te he percibido más como un hermano mayor: más sabio pero aún así cercano y enrollao. Una reunión contigo siempre me aclaró las ideas y me dió un empujón para seguir trabajando. Una grandísima parte de esta tesis te la debo a ti, no solo por el inestimable esfuerzo que has invertido en las publicaciones que la conforman, sino por la perspectiva y el aliento que me has dado siempre para hacer la parte que me tocaba a mí. Empecé la tesis sin mucho entusiasmo en la investigación por lo complicada que veía la carrera académica pero trabajar contigo me ha ayudado a descubrir qué quiero hacer y cómo hacerlo. Todo lo que escriba va a ser poco, así que solo me queda decirte que me siento tremendamente afortunado de que hayas acabado siendo mi director por "casualidad", no tanto por esta tesis, sino por lo bien que nos lo hemos pasado y lo que he disfrutado trabajando juntos. No sé cómo de factible será colaborar en el futuro o concurrir a convocatorias juntos pero espero que de vez en cuando podamos sacar un ratillo para hablar de fútbol o arreglar el mundo. Dankeschön.

Una vez agradecidos los dos "peces gordos" de esta tesis, no me puedo olvidar de todos mis compañeros de grupo: Pepe, Ángel, Stavros, Nicolas, Chema y Ricardo. Tampoco quiero olvidar a mis compañeras del grupo de Juan, especialmente Belén y Yaiza con quien más he interactuado. Cabe mencionar que Ángel fue mi primer profesor de Genética y que me avisó de que la Bioinformática podría gustarme. Ni caso te hice aunque muchas gracias por intentarlo. Muy agradecido también a Pepe por lo que he aprendido de él como profesor, como persona y también por admitirme como alumno interno en mis primeros pinitos en la Bioinfo.

Hay alguien que merece una dedicatoria especial en este apartado. Además de compañera de grupo, somos amigos desde los años del grado en Bioquímica. No contenta con dejarme sus apuntes cuando "no encontraba" los míos y avisarme cuando me olvidaba de algún trabajo o examen (cáspita, teníamos muchos), algún cuatrimestre incluso alteró su ruta a la Facultad por las mañanas para asegurarse de que asistía a clase. Gracias a ella me matriculé en Biocomputación y acabé descubriendo todo un mundo de posibilidades. En cierto modo fue gracias a ella que aprendí a programar, conocí a Pepe y a Michael y acabé yendo a Copenhague (una etapa espectacular). Para rematar la faena, su ayuda en estas últimas semanas ha sido fundamental tanto a nivel anímico como abusando de su confianza y experiencia con LaTeX para maquetar la tesis. Aunque no pueda devolverte tanto ya sabes que puedes contar conmigo para lo que quieras. Gracias Cris, por ser tan buena amiga.

También hay personas que han sido fundamentales para que lograra empezar una tesis en primer lugar. Quiero agradecer a Jacqueline por

ponerme en contacto con Juan y que me informara de las convocatorias a las que podía optar ¡muchas gracias! A todos mis profesores del instituto y a algunos del grado os agradezco vuestra dedicación. Also to you, Ruth, I learned a lot and it was always fun!

Staying at other labs during one's thesis can be challenging and sometimes stressful. However, it's normally worth it since you learn a lot and meet great people. Thanks to all my friends in Amsterdam for being so much fun: Chantal, Asli, Kulsoom, Cyrillo, Tim and Mo. I also want to thank my colleagues in Saabrucken: Tobias and Fabian, we couldn't do much outside of campus because of COVID but I learned a lot from you in those months, excited to continue! To my friends in Michigan: I had an incredible time! I learned a lot in my time in the US and I had so many unique experiences. I hope I can make it back at some point but I want to thank all of you in the meantime, especially Brian, Kaleo and Lori.

I can't forget to thank Danijela Koppers-Lalic for hosting me at the CCA, Lana Garmire for doing so at her lab in the University of Michigan and Andreas Keller at the Chair for Clinical Bioinformatics (the latter two during the pandemic). Double thanks to Andreas for his continuous advice, being so accommodating with the particular terms of current times and giving me the opportunity to continue working in his group after my thesis. Exciting times ahead!

También quiero agradecer a los amigos que conocí en Copenhague que siempre tengo ahí, aunque podamos vernos poco: Xabi, Juan, Cecilia, Joe. Ahora que al fin deposité la tesis ¡seguro que tengo tiempo de sobra para una escapada! A Migue y a Eva os agradezco estar siempre ahí para desa-

hogarme y hablar de algo que no sea investigación. A veces también está bien.

Mi familia ha sido muy importante también en este proceso, ellos han sufrido muchas veces mi ausencia por estar en el extranjero o "acabando" algún artículo. Además, siempre se han alegrado con mis logros. Agradezco a mis abuelos, tíos y primos el interés que siempre han mostrado en que me fuera bien. Un agradecimiento especial a mi tía Tatina que desde pequeño me inculcó curiosidad por la Bioquímica. Quiero incluir además a mi familia "extendida" por su comprensión y apoyo en todos los sentidos, gracias Carlos, Elena y Carlillos. A todos vosotros: ¡sois geniales!

La ambición siempre es importante pero necesita soporte y moldeamiento. Gracias Papá por animarme siempre a hacer lo que me gusta y por apoyarme emocional y económicamente cuando lo he necesitado. Durante gran parte de mi tesis fuimos "compañeros de piso" así que tuviste que lidiar con mis horarios y cambios de humor. Gracias por insistir siempre en la importancia de mi educación en todos los aspectos. Me ha costado hacerte caso pero quiero que sepas que lo valoro inmensamente. También quiero agradecer a mi hermana ser un ejemplo de excelencia y esfuerzo. Por ser yo el mayor debería ser al revés pero nos ha tocado así. Gracias siempre por tu apoyo, alegría y por echarme una mano para encajar todo cuando no estoy en España. Os quiero mucho a los dos.

Y por último, quiero agradecer a la persona que más me ha aguantado en esos momentos en los que nada parecía salir bien, en la desesperación y en la incertidumbre. A quien más ha sufrido que me quedara trabajando hasta tarde o un fin de semana: planes cancelados o que nunca se hicieron.

De no ser por ti no sé si habría llegado. Sabes que en mi lista de prioridades estás mucho más arriba que la Ciencia y el trabajo, perdóname si a veces no lo parece. No tengo palabras para decirte lo afortunado que me siento de haber vivido esta aventura (y las que nos quedan) contigo. Juntos los retos parecen menos y las ganas se multiplican. Gracias por tu cariño, tu fuerza y tu comprensión. Te quiero.

.

Bibliography

[1] A. Schmutz, C. Salignat, D. Plotkina, A. Devouassoux, T. Lee, M. Arnold, M. Ervik, and O. Kelm, "Mapping the global cancer research funding landscape," *JNCI Cancer Spectrum*, vol. 3, no. 4, 2019.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, pp. 209–249, may 2021.

[3] American Cancer Society, *The Cancer Atlas, 3rd edition.* Atlanta, 2019.

[4] M. M. Fidler, I. Soerjomataram, and F. Bray, "A global view on cancer incidence and national levels of the human development index," *International Journal of Cancer*, vol. 139, no. 11, pp. 2436–2446, 2016.

[5] R. Luengo-Fernandez, J. Leal, A. Gray, and R. Sullivan, "Economic burden of cancer across the European Union: A population-based cost analysis," *The Lancet Oncology*, vol. 14, pp. 1165–1174, nov 2013.

[6] "Novel Drug Approvals for 2019 — FDA," 2019.

[7] Evaluate Ltd., "WORLD PREVIEW 2021 Outlook to 2026...," *EvaluatePharma*, vol. 14 Edition, no. July, 2021.

[8] U. Sahin, K. Karikó, and Ö. Türeci, "MRNA-based therapeutics-developing a new class of drugs," *Nature Reviews Drug Discovery*, vol. 13, pp. 759–780, sep 2014.

[9] A. Heiser, D. Coleman, J. Dannull, D. Yancey, M. A. Maurice, C. D. Lallas, P. Dahm, D. Niedzwiecki, E. Gilboa, and J. Vieweg, "Autologous dendritic cells transfected with prostate-specific antigen RNA stimulate CTL responses against metastatic prostate tumors," *The Journal of Clinical Investigation*, vol. 109, pp. 409–417, feb 2002.

[10] B. Weide, J. P. Carralot, A. Reese, B. Scheel, T. K. Eigentler, I. Hoerr, H. G. Rammensee, C. Garbe, and S. Pascolowz, "Results of the first phase I/II clinical vaccination trial with direct injection of mRNA," *Journal of Immunotherapy*, vol. 31, pp. 180–188, feb 2008.

[11] K.Barnes, "The first monoclonal antibody therapy," *Nature*, p. 2, 2018.

[12] J. Li and Z. Zhu, "Research and development of next generation of antibody-based therapeutics," *Acta Pharmacologica Sinica 2010 31:9*, vol. 31, no. 9, pp. 1198–1207, 2010.

[13] J. M. Reichert, "Antibody therapeutics approved or in regulatory review in the EU or US," 2021.

[14] A. Mullard, "FDA approves 100th monoclonal antibody product," jul 2021.

[15] J. K. Liu, "The history of monoclonal antibody development - Progress, remaining challenges and future innovations," *Annals of Medicine and Surgery*, vol. 3, no. 4, pp. 113–116, 2014.

[16] J. M. Unger, E. Cook, E. Tai, and A. Bleyer, "The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies," *American Society of Clinical Oncology Educational Book*, vol. 36, pp. 185–198, jun 2016.

[17] J. M. Unger, D. L. Hershman, C. Till, L. M. Minasian, R. U. Osarogiagbon, M. E. Fleury, and R. Vaidya, ""When Offered to Participate": A Systematic Review and Meta-Analysis of Patient Agreement to Participate in Cancer Clinical Trials," *Journal of the National Cancer Institute*, vol. 113, pp. 244–257, mar 2021.

[18] "Budget Appropriation for Fiscal Year 2020 - NIH: National Institute of Allergy and Infectious Diseases."

[19] "Chan Zuckerberg Initiative.."

[20] "Bill & Melinda Gates Foundation."

[21] "Zara founder to spend $344 million on breast cancer-screening for Spanish hospitals - Reuters."

[22] R. Sengupta and K. Honey, "Aacr cancer progress report 2020: Turning science into lifesaving care," *Clinical Cancer Research*, vol. 26, no. 19, pp. 5055–5055, 2020.

[23] M. López-Gómez, E. Malmierca, M. de Górgolas, and E. Casado, "Cancer in developing countries: The next most preventable pan-

demic. The global problem of cancer," *Critical Reviews in Oncology/Hematology*, vol. 88, pp. 117–122, oct 2013.

[24] J. Baselga, N. Bhardwaj, L. C. Cantley, R. DeMatteo, R. N. DuBois, M. Foti, S. M. Gapstur, W. C. Hahn, L. J. Helman, R. A. Jensen, E. D. Paskett, T. S. Lawrence, S. G. Lutzker, and E. Szabo, "AACR Cancer Progress Report 2015," 2015.

[25] World Health Organization, "Cancer," 2021.

[26] "What Is Cancer? - National Cancer Institute," 2021.

[27] M. Quaresma, M. P. Coleman, and B. Rachet, "40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: A population-based study," *The Lancet*, vol. 385, pp. 1206–1218, mar 2015.

[28] "Lung Cancer - Non-Small Cell: Statistics — Cancer.Net," 2021.

[29] H. Dillekås, M. S. Rogers, and O. Straume, "Are 90% of deaths from cancer caused by metastases?," *Cancer Medicine*, vol. 8, pp. 5574–5576, sep 2019.

[30] D. Hanahan and R. A. Weinberg, "The Hallmarks of Cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[31] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, pp. 646–674, mar 2011.

[32] T. Estape, "Cancer in the elderly: Challenges and barriers," *Asia-Pacific Journal of Oncology Nursing*, vol. 5, no. 1, pp. 40–42, 2018.

[33] G. G. Chen and P. B. Lai, *Apoptosis in carcinogenesis and chemotherapy: Apoptosis in cancer*. Springer Netherlands, 2009.

[34] C. Wang and R. J. Youle, "The Role of Mitochondria in Apoptosis," *Annual review of genetics*, vol. 43, p. 95, 2009.

[35] P. E. D. Costa, "Robbins' pathologic basis of disease. R. S. Cotran, V. Kumar and S. L. Robbins. W. B. Saunders, Philadelphia, 1989. No. of pages: 1519. Price £37. ISBN:0 7216 2302 6," *The Journal of Pathology*, vol. 160, no. 1, p. 89, 1990.

[36] P. Jaime-Sanchez, I. Uranga-Murillo, N. Aguilo, S. C. Khouili, M. A. Arias, D. Sancho, and J. Pardo, "Cell death induced by cytotoxic CD8 + T cells is immunogenic and primes caspase-3-dependent spread immunity against endogenous tumor antigens," *Journal for ImmunoTherapy of Cancer*, vol. 8, no. 1, p. e000528, 2020.

[37] Leslie A. Pray, "Errors in DNA Replication — Learn Science at Scitable," 2008.

[38] B. N. Ames, M. K. Shigenaga, and T. M. Hagen, "Oxidants, antioxidants, and the degenerative diseases of aging," 1993.

[39] J. Fowles and E. Dybing, "Application of toxicological risk assessment principles to the chemical constituents of cigarette smoke," *Tobacco Control*, vol. 12, no. 4, p. 424, 2003.

[40] A. J. Sasco, M. B. Secretan, and K. Straif, "Tobacco smoking and cancer: A brief review of recent epidemiological evidence," *Lung Cancer*, vol. 45, no. SUPPL. 2, pp. S3—-S9, 2004.

[41] A. Weston and C. C. Harris, "Chemical Carcinogenesis," in *Holland-Frei Cancer Medicine* (et al. ufe DW, Pollock RE, Weichselbaum RR, ed.), ch. Chapter 17, Hamilton (ON): BC Decker; 2003., 6 ed., 2003.

[42] J. E. Cleaver and D. L. Mitchell, *Ultraviolet Radiation Carcinogenesis.* BC Decker, 2000.

[43] L. R. Ferguson, "Meat and cancer," *Meat Science*, vol. 84, pp. 308–313, feb 2010.

[44] V. Bagnardi, M. Blangiardo, C. L. Vecchia, and G. Corrao, "Alcohol Consumption and the Risk of Cancer: A Meta-Analysis," *Alcohol Research & Health*, vol. 25, no. 4, p. 263, 2001.

[45] G. Obe and H. Ristow, "Mutagenic, cancerogenic and teratogenic effects of alcohol," *Mutation Research/Reviews in Genetic Toxicology*, vol. 65, no. 4, pp. 229–259, 1979.

[46] K. Kabsch and A. Alonso, "The human papillomavirus type 16 (HPV-16) E5 protein sensitizes human keratinocytes to apoptosis induced by osmotic stress," *Oncogene*, vol. 21, pp. 947–953, feb 2002.

[47] A. Mantovani, "Molecular Pathways Linking Inflammation and Cancer," *Current Molecular Medicine*, vol. 10, pp. 369–373, may 2010.

[48] K. Taniguchi and M. Karin, "IL-6 and related cytokines as the critical lynchpins between inflammation and cancer," *Seminars in Immunology*, vol. 26, pp. 54–74, feb 2014.

[49] W. KY, C. K, and C. GA, "Obesity and cancer," *The oncologist*, vol. 15, no. 6, pp. 197–204, 2010.

[50] J. Clague and L. Bernstein, "Physical activity and cancer," *Current Oncology Reports*, vol. 14, no. 6, pp. 550–558, 2012.

[51] H. Vainio, R. Kaaks, and F. Bianchini, "Weight control and physical activity in cancer prevention: International evaluation of the evidence," *European Journal of Cancer Prevention*, vol. 11, no. SUPPL. 2, pp. S94—-100, 2002.

[52] C. M. Friedenreich and M. R. Orenstein, "Physical activity and cancer prevention: Etiologic evidence and biological mechanisms," *Journal of Nutrition*, vol. 132, pp. 3456S—-3464S, nov 2002.

[53] L. H. Kushi, T. Byers, C. Doyle, E. V. Bandera, M. McCullough, T. Gansler, K. S. Andrews, and M. J. Thun, "American Cancer Society Guidelines on Nutrition and Physical Activity for Cancer Prevention: Reducing the Risk of Cancer With Healthy Food Choices and Physical Activity," *CA: A Cancer Journal for Clinicians*, vol. 56, pp. 254–281, sep 2006.

[54] B. L. Henderson BE, "Hormones and the Etiology of Cancer," in *Holland-Frei Cancer Medicine*, BC Decker, 2000.

[55] B. E. Henderson, R. K. Ross, M. C. Pike, and J. T. Casagrande, "Endogenous Hormones as a Major Factor in Human Cancer," *Cancer Research*, vol. 42, no. 8, pp. 3232–3239, 1982.

[56] B. E. Henderson, R. Ross, and L. Bernstein, "Estrogens as a Cause of Human Cancer: The Richard and Hinda Rosenthal Foundation Award Lecture," *Cancer Research*, vol. 48, no. 2, 1988.

[57] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki, "Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland," *http://dx.doi.org/10.1056/NEJM200007133430201*, vol. 343, no. 2, pp. 78–85, 2009.

[58] V. Fanfani, L. Citi, A. L. Harris, F. Pezzella, and G. Stracquadanio, "The landscape of the heritable cancer genome," *Cancer Research*, vol. 81, pp. 2588–2599, may 2021.

[59] C. K, L. P, and H. K, "Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database," *International journal of cancer*, vol. 99, pp. 260–266, may 2002.

[60] J. Yokota, "Tumor progression and metastasis," *Carcinogenesis*, vol. 21, pp. 497–503, mar 2000.

[61] Editorial, "Tumor Progression," *Cancer Research*, vol. 17, no. 5, 1957.

[62] P. J. Fialkow, "Clonal origin of human tumors.," nov 1979.

[63] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[64] M. C. van den Berg, L. MacCarthy-Morrogh, D. Carter, J. Morris, I. Ribeiro Bravo, Y. Feng, and P. Martin, "Proteolytic and Opportunistic Breaching of the Basement Membrane Zone by Immune Cells

during Tumor Initiation," *Cell Reports*, vol. 27, pp. 2837–2846.e4, jun 2019.

[65] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[66] I. Carr, "Lymphatic metastasis," *CANCER AND METASTASIS REVIEW*, vol. 2, no. 3, pp. 307–317, 1983.

[67] I. R. Beavon, "The E-cadherin-catenin complex in tumour metastasisstructure, function and regulation," *European Journal of Cancer*, vol. 36, no. 13, pp. 1607–1620, 2000.

[68] K. Pantel and M. R. Speicher, "The biology of circulating tumor cells," *Oncogene*, vol. 35, pp. 1216–1224, mar 2016.

[69] J. Fares, M. Y. Fares, H. H. Khachfe, H. A. Salhab, and Y. Fares, "Molecular principles of metastasis: a hallmark of cancer revisited," *Signal Transduction and Targeted Therapy*, vol. 5, pp. 1–17, mar 2020.

[70] S. Y. Wong and R. O. Hynes, "Lymphatic or Hematogenous Dissemination: How Does a Metastatic Tumor Cell Decide?," *Cell cycle (Georgetown, Tex.)*, vol. 5, no. 8, p. 812, 2006.

[71] L. Bubendorf, A. Schöpfer, U. Wagner, G. Sauter, H. Moch, N. Willi, T. C. Gasser, and M. J. Mihatsch, "Metastatic patterns of prostate cancer: An autopsy study of 1,589 patients," *Human Pathology*, vol. 31, no. 5, pp. 578–583, 2000.

[72] S. Paget, "THE DISTRIBUTION OF SECONDARY GROWTHS IN CANCER OF THE BREAST.," *The Lancet*, vol. 133, pp. 571–573, mar 1889.

[73] R. R. Langley and I. J. Fidler, "The seed and soil hypothesis revisited-The role of tumor-stroma interactions in metastasis to different organs," *International Journal of Cancer*, vol. 128, pp. 2527–2535, jun 2011.

[74] C. A. Klein, "Parallel progression of primary tumours and metastases," *Nature Reviews Cancer 2009 9:4*, vol. 9, no. 4, pp. 302–312, 2009.

[75] G. G. Steel and L. F. Lamerton, "The growth rate of human tumours," *British Journal of Cancer*, vol. 20, no. 1, pp. 74–86, 1966.

[76] V. P. Collins, R. K. Loeffler, and H. Tivey, "Observations on growth rates of human tumors," *The American journal of roentgenology, radium therapy, and nuclear medicine*, vol. 76, no. 5, pp. 988–1000, 1956.

[77] H. Yoo, B. H. Nam, H. S. Yang, H. S. Sang, S. L. Jin, and H. L. Seung, "Growth rates of metastatic brain tumors in nonsmall cell lung cancer," *Cancer*, vol. 113, pp. 1043–1047, sep 2008.

[78] H. R. Withers and S. P. Lee, "Modeling growth kinetics and statistical distribution of oligometastases," *Seminars in Radiation Oncology*, vol. 16, no. 2, pp. 111–119, 2006.

[79] J. L. Abbruzzese, M. C. Abbruzzese, K. R. Hess, M. N. Raber, R. Lenzi, and P. Frost, "Unknown primary carcinoma: Natural history and prognostic factors in 657 consecutive patients," *Journal of Clinical Oncology*, vol. 12, no. 6, pp. 1272–1280, 1994.

[80] A. J. Van de Wouw, M. L. Janssen-Heijnen, J. W. Coebergh, and H. F. Hillen, "Epidemiology of unknown primary tumours; incidence and population-based survival of 1285 patients in Southeast Netherlands, 1984-1992," *European Journal of Cancer*, vol. 38, no. 3, pp. 409–413, 2002.

[81] E. J, E. R, K. J, S. M, F. G, R. R, S. H, S. HJ, and H. D, "The process of metastasisation for breast cancer," *European journal of cancer (Oxford, England : 1990)*, vol. 39, no. 12, pp. 1794–1806, 2003.

[82] S. B. Edge and C. C. Compton, "The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM," *Annals of Surgical Oncology 2010 17:6*, vol. 17, pp. 1471–1474, feb 2010.

[83] B. Spiessl, O. H. Beahrs, and P. Hermanek, "TNM Atlas," *Anti-Cancer Drugs*, vol. 1, no. 1, p. 89, 1990.

[84] J. T. Loud and J. Murphy, "Cancer Screening and Early Detection in the 21st Century," *Seminars in Oncology Nursing*, vol. 33, pp. 121–128, may 2017.

[85] S. McPhail, S. Johnson, D. Greenberg, M. Peake, and B. Rous, "Stage at diagnosis and early mortality from cancer in England," *British Journal of Cancer*, vol. 112, pp. S108–S115, mar 2015.

[86] M. Sant, C. Allemani, R. Capocaccia, T. Hakulinen, T. Aareleid, J. W. Coebergh, M. P. Coleman, P. Grosclaude, C. Martinez, J. Bell, J. Youngson, F. Berrino, A. Kupp, G. Hedelin, G. Chaplain, C. Exbrayat, B. Tretarre, J. Mace-Lesech, A. Danzon, M. Mercier, N. Raverdy, E. Artioli, M. Federico, A. Barchielli, E. Paci, G. Gatta, P. Crosignani, D. Speciale, M. R. Ruzza, E. Frassoldi, A. Verdecchia, L. Gafa, R. Tumino, M. La Rosa, A. Voogd, and E. M. Williams, "Stage at diagnosis is a key explanation of differences in breast cancer survival across Europe," *International Journal of Cancer*, vol. 106, pp. 416–422, sep 2003.

[87] T. Byers, R. C. Wender, A. Jemal, A. M. Baskies, E. E. Ward, and O. W. Brawley, "The American Cancer Society challenge goal to reduce US cancer mortality by 50% between 1990 and 2015: Results and reflections," *CA: A Cancer Journal for Clinicians*, vol. 66, pp. 359–369, sep 2016.

[88] Z. Kakushadze, R. Raghubanshi, and W. Yu, "Estimating cost savings from early cancer diagnosis," *Data*, vol. 2, p. 30, sep 2017.

[89] J. D. Schiffman, P. G. Fisher, and P. Gibbs, "Early Detection of Cancer: Past, Present, and Future," *American Society of Clinical Oncology Educational Book*, pp. 57–65, may 2015.

[90] "Guardant Health – Conquering Cancer with Data — Guardant Health."

[91] A. Diamantis, E. Magiorkinis, and H. Koutselini, "Fine-needle aspiration (FNA) biopsy: Historical aspects," 2009.

[92] E. Crowley, F. Di Nicolantonio, F. Loupakis, and A. Bardelli, "Liquid biopsy: Monitoring cancer-genetics in the blood," *Nature Reviews Clinical Oncology*, vol. 10, pp. 472–484, jul 2013.

[93] S. Misale, R. Yaeger, S. Hobor, E. Scala, M. Janakiraman, D. Liska, E. Valtorta, R. Schiavo, M. Buscarino, G. Siravegna, K. Bencardino, A. Cercek, C. T. Chen, S. Veronese, C. Zanon, A. Sartore-Bianchi, M. Gambacorta, M. Gallicchio, E. Vakiani, V. Boscaro, E. Medico, M. Weiser, S. Siena, F. Di Nicolantonio, D. Solit, and A. Bardelli, "Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer," *Nature*, vol. 486, pp. 532–536, jun 2012.

[94] L. A. Diaz, R. T. Williams, J. Wu, I. Kinde, J. R. Hecht, J. Berlin, B. Allen, I. Bozic, J. G. Reiter, M. A. Nowak, K. W. Kinzler, K. S. Oliner, and B. Vogelstein, "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers," *Nature*, vol. 486, pp. 537–540, jun 2012.

[95] E. Aparicio-Puerta, D. Jáspez, R. Lebrón, D. Koppers-Lalic, J. A. Marchal, and M. Hackenberg, "IiqDB: A small-RNAseq knowledge discovery database for liquid biopsy studies," *Nucleic Acids Research*, vol. 47, no. D1, pp. D113–D120, 2019.

[96] G. De Rubis, S. Rajeev Krishnan, and M. Bebawy, "Liquid biopsies in cancer diagnosis, monitoring, and prognosis," *Trends in Pharmacological Sciences*, vol. 40, no. 3, pp. 172–186, 2019.

[97] P. P. Praharaj, S. K. Bhutia, S. Nagrath, R. L. Bitting, and G. Deep, "Circulating tumor cell-derived organoids: Current challenges and promises in medical research and precision medicine," *Biochimica et Biophysica Acta - Reviews on Cancer*, vol. 1869, no. 2, pp. 117–127, 2018.

[98] S. Sharma, R. Zhuang, M. Long, M. Pavlovic, Y. Kang, A. Ilyas, and W. Asghar, "Circulating tumor cell isolation, culture, and downstream molecular analysis," *Biotechnology Advances*, vol. 36, pp. 1063–1078, jul 2018.

[99] J. T. Kaifi, M. Kunkel, D. T. Dicker, J. Joude, J. E. Allen, A. Das, J. Zhu, Z. Yang, N. E. Sarwani, G. Li, K. F. Staveley-O'Carroll, and W. S. El-Deiry, "Circulating tumor cell levels are elevated in colorectal cancer patients with high tumor burden in the liver," *Cancer Biology and Therapy*, vol. 16, no. 5, pp. 690–698, 2015.

[100] Y. Mishima, B. Paiva, J. Shi, J. Park, S. Manier, S. Takagi, M. Massoud, A. Perilla-Glen, Y. Aljawai, D. Huynh, A. M. Roccaro, A. Sacco, M. Capelletti, A. Detappe, D. Alignani, K. C. Anderson, N. C. Munshi, F. Prosper, J. G. Lohr, G. Ha, S. S. Freeman, E. M. Van Allen, V. A. Adalsteinsson, F. Michor, J. F. San Miguel, and I. M. Ghobrial, "The Mutational Landscape of Circulating Tumor Cells in Multiple Myeloma," *Cell Reports*, vol. 19, no. 1, pp. 218–224, 2017.

[101] A. Lyberopoulou, G. Aravantinos, E. P. Efstathopoulos, N. Nikiteas, P. Bouziotis, A. Isaakidou, A. Papalois, E. Marinos, and M. Gazouli, "Mutational analysis of circulating tumor cells from colorectal cancer patients and correlation with primary tumor tissue," *PLoS ONE*, vol. 10, no. 4, 2015.

[102] M. Giuliano, A. Giordano, S. Jackson, K. R. Hess, U. De Giorgi, M. Mego, B. C. Handy, N. T. Ueno, R. H. Alvarez, M. De Laurentiis, S. De Placido, V. Valero, G. N. Hortobagyi, J. M. Reuben, and M. Cristofanilli, "Circulating tumor cells as prognostic and predictive markers in metastatic breast cancer patients receiving first-line systemic treatment," *Breast Cancer Research*, vol. 13, jun 2011.

[103] S. Maheswaran, L. V. Sequist, S. Nagrath, L. Ulkus, B. Brannigan, C. V. Collura, E. Inserra, S. Diederichs, A. J. Iafrate, D. W. Bell, S. Digumarthy, A. Muzikansky, D. Irimia, J. Settleman, R. G. Tompkins, T. J. Lynch, M. Toner, and D. A. Haber, "Detection of mutations in EGFR in circulating lung-cancer cells," *The New England journal of medicine*, vol. 359, pp. 366–377, jul 2008.

[104] S. L. Stott, L. Richard, S. Nagrath, Y. Min, D. T. Miyamoto, L. Ulkus, E. J. Inserra, M. Ulman, S. Springer, Z. Nakamura, A. L. Moore, D. Tsukrov, M. E. Kempner, D. M. Dahl, W. Chin-lee, J. A. Iafrate, M. R. Smith, R. G. Tompkins, L. V. Sequist, M. Toner, D. A. Haber, and S. Maheswaran, "Isolation and characterization of circulating tumor cells from patients with localized and metastatic prostate cancer," *Science Translational Medicine*, vol. 2, no. 25, 2010.

[105] M. C. Liu, P. G. Shields, R. D. Warren, P. Cohen, M. Wilkinson, Y. L. Ottaviano, S. B. Rao, J. Eng-Wong, F. Seillier-Moiseiwitsch, A. M. Noone, and C. Isaacs, "Circulating tumor cells: A useful predictor of treatment efficacy in metastatic breast cancer," *Journal of Clinical Oncology*, vol. 27, pp. 5153–5159, nov 2009.

[106] D. F. Hayes, M. Cristofanilli, G. T. Budd, M. J. Ellis, A. Stopeck, M. C. Miller, J. Matera, W. J. Allard, G. V. Doyle, and L. W. W. M. Terstappen, "Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 12, no. 14 Pt 1, pp. 4218–4224, 2006.

[107] A. J. Rushton, G. Nteliopoulos, J. A. Shaw, and R. C. Coombes, "A review of circulating tumour cell enrichment technologies," *Cancers*, vol. 13, pp. 1–33, mar 2021.

[108] T. Sato, D. E. Stange, M. Ferrante, R. G. Vries, J. H. Van Es, S. Van Den Brink, W. J. Van Houdt, A. Pronk, J. Van Gorp, P. D. Siersema, and H. Clevers, "Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium," *Gastroenterology*, vol. 141, no. 5, pp. 1762–1772, 2011.

[109] D. Gao, I. Vela, A. Sboner, P. J. Iaquinta, W. R. Karthaus, A. Gopalan, C. Dowling, J. N. Wanjala, E. A. Undvall, V. K. Arora, J. Wongvipat, M. Kossai, S. Ramazanoglu, L. P. Barboza, W. Di, Z. Cao, Q. F. Zhang, I. Sirota, L. Ran, T. Y. Macdonald, H. Beltran,

J. M. Mosquera, K. A. Touijer, P. T. Scardino, V. P. Laudone, K. R. Curtis, D. E. Rathkopf, M. J. Morris, D. C. Danila, S. F. Slovin, S. B. Solomon, J. A. Eastham, P. Chi, B. Carver, M. A. Rubin, H. I. Scher, H. Clevers, C. L. Sawyers, and Y. Chen, "Organoid cultures derived from patients with advanced prostate cancer," *Cell*, vol. 159, pp. 176–187, sep 2014.

[110] E. M. Veziroglu and G. I. Mias, "Characterizing Extracellular Vesicles and Their Diverse RNA Contents," *Frontiers in Genetics*, vol. 11, jul 2020.

[111] B. J. Crenshaw, L. Gu, B. Sims, and Q. L. Matthews, "Exosome Biogenesis and Biological Function in Response to Viral Infections," *The Open Virology Journal*, vol. 12, no. 1, pp. 134–148, 2018.

[112] R. Kalluri and V. S. LeBleu, "The biology, function, and biomedical applications of exosomes," *Science*, vol. 367, no. 6478, 2020.

[113] B. Zhou, K. Xu, X. Zheng, T. Chen, J. Wang, Y. Song, Y. Shao, and S. Zheng, "Application of exosomes as liquid biopsy in clinical diagnosis," *Signal Transduction and Targeted Therapy*, vol. 5, no. 1, pp. 1–14, 2020.

[114] R. Drula, L. F. Ott, I. Berindan-Neagoe, K. Pantel, and G. A. Calin, "Micrornas from liquid biopsy derived extracellular vesicles: Recent advances in detection and characterization methods," *Cancers*, vol. 12, no. 8, pp. 1–24, 2020.

[115] D. Bhagirath, T. L. Yang, N. Bucay, K. Sekhon, S. Majid, V. Shahryari, R. Dahiya, Y. Tanaka, and S. Saini, "microRNA-1246

is an exosomal biomarker for aggressive prostate cancer," *Cancer Research*, vol. 78, no. 7, pp. 1833–1844, 2018.

[116] S. Ebrahimkhani, F. Vafaee, S. Hallal, H. Wei, M. Y. T. Lee, P. E. Young, L. Satgunaseelan, H. Beadnall, M. H. Barnett, B. Shivalingam, C. M. Suter, M. E. Buckland, and K. L. Kaufman, "Deep sequencing of circulating exosomal microRNA allows non-invasive glioblastoma diagnosis," *npj Precision Oncology*, vol. 2, no. 1, 2018.

[117] E. Castellanos-Rizaldos, X. Zhang, V. R. Tadigotla, D. G. Grimm, C. Karlovich, L. E. Raez, and J. K. Skog, "Exosome-based detection of activating and resistance EGFR mutations from plasma of non-small cell lung cancer patients," *Oncotarget*, vol. 10, no. 30, pp. 2911–2920, 2019.

[118] M. Colletti, A. Paolini, A. Galardi, V. Di Paolo, L. Pascucci, I. Russo, B. De Angelis, H. Peinado, R. De Vito, G. M. Milano, F. Locatelli, A. Masotti, and A. Di Giannatale, "Expression profiles of exosomal miRNAs isolated from plasma of patients with desmoplastic small round cell tumor," *Epigenomics*, vol. 11, no. 5, pp. 489–500, 2019.

[119] R. Shi, P. Y. Wang, X. Y. Li, J. X. Chen, Y. Li, X. Z. Zhang, C. G. Zhang, T. Jiang, W. B. Li, W. Ding, and S. J. Cheng, "Exosomal levels of miRNA-21 from cerebrospinal fluids associated with poor prognosis and tumor recurrence of glioma patients," *Oncotarget*, vol. 6, no. 29, pp. 26971–26981, 2015.

[120] M. D. Giráldez, J. J. Lozano, G. Ramírez, E. Hijona, L. Bujanda, A. Castells, and M. Gironella, "Circulating MicroRNAs as biomark-

ers of colorectal cancer: Results from a genome-wide profiling and validation study," *Clinical Gastroenterology and Hepatology*, vol. 11, no. 6, 2013.

[121] N. Hayashi, A. Yamasaki, S. Ueda, S. Okazaki, Y. Ohno, T. Tanaka, Y. Endo, Y. Tomioka, K. Masuko, T. Masuko, and R. Sugiura, "Oncogenic transformation of NIH/3T3 cells by the overexpression of L-type amino acid transporter 1, a promising anti-cancer target," *Oncotarget*, vol. 12, pp. 1256–1270, jun 2021.

[122] R. J. Leary, M. Sausen, I. Kinde, N. Papadopoulos, J. D. Carpten, D. Craig, J. O'Shaughnessy, K. W. Kinzler, G. Parmigiani, B. Vogelstein, L. A. Diaz, and V. E. Velculescu, "Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing," *Science Translational Medicine*, vol. 4, no. 162, 2012.

[123] E. Heitzer, M. Auer, E. M. Hoffmann, M. Pichler, C. Gasch, P. Ulz, S. Lax, J. Waldispuehl-Geigl, O. Mauermann, S. Mohan, G. Pristauz, C. Lackner, G. Höfler, F. Eisner, E. Petru, H. Sill, H. Samonigg, K. Pantel, S. Riethdorf, T. Bauernhofer, J. B. Geigl, and M. R. Speicher, "Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer," *International Journal of Cancer*, vol. 133, pp. 346–356, jul 2013.

[124] A. R. Thierry, F. Mouliere, S. El Messaoudi, C. Mollevi, E. Lopez-Crapez, F. Rolet, B. Gillet, C. Gongora, P. Dechelotte, B. Robert, M. Del Rio, P. J. Lamy, F. Bibeau, M. Nouaille, V. Loriot, A. S. Jarrousse, F. Molina, M. Mathonnet, D. Pezet, and M. Ychou, "Clinical

validation of the detection of KRAS and BRAF mutations from circulating tumor DNA," *Nature Medicine*, vol. 20, no. 4, pp. 430–435, 2014.

[125] C. Bettegowda, M. Sausen, R. J. Leary, I. Kinde, Y. Wang, N. Agrawal, B. R. Bartlett, H. Wang, B. Luber, R. M. Alani, E. S. Antonarakis, N. S. Azad, A. Bardelli, H. Brem, J. L. Cameron, C. C. Lee, L. A. Fecher, G. L. Gallia, P. Gibbs, D. Le, R. L. Giuntoli, M. Goggins, M. D. Hogarty, M. Holdhoff, S. M. Hong, Y. Jiao, H. H. Juhl, J. J. Kim, G. Siravegna, D. A. Laheru, C. Lauricella, M. Lim, E. J. Lipson, S. K. N. Marie, G. J. Netto, K. S. Oliner, A. Olivi, L. Olsson, G. J. Riggins, A. Sartore-Bianchi, K. Schmidt, I. M. Shih, S. M. Oba-Shinjo, S. Siena, D. Theodorescu, J. Tie, T. T. Harkins, S. Veronese, T. L. Wang, J. D. Weingart, C. L. Wolfgang, L. D. Wood, D. Xing, R. H. Hruban, J. Wu, P. J. Allen, C. M. Schmidt, M. A. Choti, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, N. Papadopoulos, and L. A. Diaz, "Detection of circulating tumor DNA in early- and late-stage human malignancies," *Science Translational Medicine*, vol. 6, feb 2014.

[126] E. Gormally, E. Caboux, P. Vineis, and P. Hainaut, "Circulating free DNA in plasma or serum as biomarker of carcinogenesis: Practical aspects and biological significance," *Mutation Research - Reviews in Mutation Research*, vol. 635, pp. 105–117, may 2007.

[127] P. O. Delgado, B. C. A. Alves, F. De Sousa Gehrke, R. K. Kuniyoshi, M. L. Wroclavski, A. Del Giglio, and F. L. A. Fonseca, "Characteriza-

tion of cell-free circulating DNA in plasma in patients with prostate cancer," *Tumor Biology*, vol. 34, no. 2, pp. 983–986, 2013.

[128] M. Stroun, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Beljanski, "Neoplastic characteristics of the DNA found in the plasma of cancer patients," *Oncology*, vol. 46, no. 5, pp. 318–322, 1989.

[129] A. R. Thierry, F. Mouliere, C. Gongora, J. Ollier, B. Robert, M. Ychou, M. del Rio, and F. Molina, "Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts," *Nucleic acids research*, vol. 38, pp. 6159–6175, may 2010.

[130] K. Kim, D. G. Shin, M. K. Park, S. H. Baik, T. H. Kim, S. Kim, and S. Y. Lee, "Circulating cell-free DNA as a promising biomarker in patients with gastric cancer: Diagnostic validity and significant reduction of cfDNA after surgical resection," *Annals of Surgical Treatment and Research*, vol. 86, no. 3, pp. 136–142, 2014.

[131] G. Vandekerkhove, J. M. Lavoie, M. Annala, A. J. Murtha, N. Sundahl, S. Walz, T. Sano, S. Taavitsainen, E. Ritch, L. Fazli, A. Hurtado-Coll, G. Wang, M. Nykter, P. C. Black, T. Todenhöfer, P. Ost, E. A. Gibb, K. N. Chi, B. J. Eigl, and A. W. Wyatt, "Plasma ctDNA is a tumor tissue surrogate and enables clinical-genomic stratification of metastatic bladder cancer," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021.

[132] T. Lecomte, A. Berger, F. Zinzindohoué, S. Micard, B. Landi, H. Blons, P. Beaune, P. H. Cugnenc, and P. Laurent-Puig, "Detection of free-circulating tumor-associated DNA in plasma of colorectal can-

cer patients and its association with prognosis," *International Journal of Cancer*, vol. 100, no. 5, pp. 542–548, 2002.

[133] A. M. Newman, S. V. Bratman, J. To, J. F. Wynne, N. C. Eclov, L. A. Modlin, C. L. Liu, J. W. Neal, H. A. Wakelee, R. E. Merritt, J. B. Shrager, B. W. Loo, A. A. Alizadeh, and M. Diehn, "An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage," *Nature Medicine*, vol. 20, no. 5, pp. 548–554, 2014.

[134] F. Diehl, K. Schmidt, M. A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, L. Sokoll, S. A. Szabo, K. W. Kinzler, B. Vogelstein, and L. A. Diaz, "Circulating mutant DNA to assess tumor dynamics," *Nature medicine*, vol. 14, pp. 985–990, sep 2008.

[135] V. Taly, D. Pekin, L. Benhaim, S. K. Kotsopoulos, D. L. Corre, X. Li, I. Atochin, D. R. Link, A. D. Griffiths, K. Pallier, H. Blons, O. Bouché, B. Landi, J. B. Hutchison, and P. Laurent-Puig, "Multiplex picodroplet digital PCR to detect KRAS mutations in circulating DNA from the plasma of colorectal cancer patients," *Clinical Chemistry*, vol. 59, no. 12, pp. 1722–1731, 2013.

[136] E. Heitzer, P. Ulz, J. Belic, S. Gutschi, F. Quehenberger, K. Fischereder, T. Benezeder, M. Auer, C. Pischler, S. Mannweiler, M. Pichler, F. Eisner, M. Haeusler, S. Riethdorf, K. Pantel, H. Samonigg, G. Hoefler, H. Augustin, J. B. Geigl, and M. R. Speicher, "Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing," *Genome Medicine*, vol. 5, no. 4, 2013.

[137] M. F. De Souza, H. Kuasne, M. D. C. Barros-Filho, H. L. Cilião, F. A. Marchi, P. E. Fuganti, A. R. Paschoal, S. R. Rogatto, and I. M. De Syllos Cólus, "Circulating mRNAs and miRNAs as candidate markers for the diagnosis and prognosis of prostate cancer," *PLoS ONE*, vol. 12, sep 2017.

[138] G. Tzimagiorgis, E. Z. Michailidou, A. Kritis, A. K. Markopoulos, and S. Kouidou, "Recovering circulating extracellular or cell-free RNA from bodily fluids," *Cancer Epidemiology*, vol. 35, no. 6, pp. 580–589, 2011.

[139] V. Armand-Labit and A. Pradines, "Circulating cell-free microRNAs as clinical cancer biomarkers," *Biomolecular Concepts*, vol. 8, pp. 61–81, may 2017.

[140] H. Zhang, D. Zhou, M. Ying, M. Chen, P. Chen, Z. Chen, and F. Zhang, "Expression of long non-coding RNA (LncRNA) small nucleolar rna host gene 1 (SNHG1) exacerbates hepatocellular carcinoma through suppressing miR-195," *Medical Science Monitor*, vol. 22, pp. 4820–4829, 2016.

[141] S. Zhong, M. C. Ng, Y. M. Lo, J. C. Chan, and P. J. Johnson, "Presence of mitochondrial tRNA(Leu(UUR)) A to G 3243 mutation in DNA extracted from serum and plasma of patients with type 2 diabetes mellitus," *Journal of Clinical Pathology*, vol. 53, no. 6, pp. 466–469, 2000.

[142] J. Qin, T. L. Williams, and M. R. Fernando, "A novel blood collection device stabilizes cell-free RNA in blood during sample shipping and storage," *BMC Research Notes*, vol. 6, no. 1, 2013.

[143] O. Pös, O. Biró, T. Szemes, and B. Nagy, "Circulating cell-free nucleic acids: Characteristics and applications," *European Journal of Human Genetics*, vol. 26, no. 7, pp. 937–945, 2018.

[144] C. L. Shih, J. D. Luo, J. W. C. Chang, T. L. Chen, Y. T. Chien, C. J. Yu, and C. C. Chiou, "Circulating messenger RNA profiling with microarray and next-generation sequencing: Cross-platform comparison," *Cancer Genomics and Proteomics*, vol. 12, no. 5, pp. 223–230, 2015.

[145] V. W. Xue, M. T. Cheung, P. T. Chan, L. L. Luk, V. H. Lee, T. C. Au, A. C. Yu, W. C. Cho, H. F. A. Tsang, A. K. Chan, and S. C. Wong, "Non-invasive Potential Circulating mRNA Markers for Colorectal Adenoma Using Targeted Sequencing," *Scientific Reports*, vol. 9, pp. 1–10, sep 2019.

[146] M. D. Giraldez, R. M. Spengler, A. Etheridge, A. J. Goicochea, M. Tuck, S. W. Choi, D. J. Galas, and M. Tewari, "Phospho-RNA-seq: a modified small RNA-seq method that reveals circulating mRNA and lncRNA fragments as potential biomarkers in human plasma," *The EMBO Journal*, vol. 38, jun 2019.

[147] J. G. Ruby, C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel, "Large-Scale Sequencing Reveals 21U-RNAs

and Additional MicroRNAs and Endogenous siRNAs in C. elegans," *Cell*, vol. 127, no. 6, pp. 1193–1207, 2006.

[148] P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin, and M. Tewari, "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 10513–10518, jul 2008.

[149] W. Tan, B. Liu, S. Qu, G. Liang, W. Luo, and C. Gong, "MicroRNAs and cancer: Key paradigms in molecular therapy (Review)," *Oncology Letters*, vol. 15, pp. 2735–2742, mar 2018.

[150] A. Navarro, T. Díaz, N. Tovar, F. Pedrosa, R. Tejero, M. T. Cibeira, L. Magnano, L. Rosiñol, M. Monzó, J. Bladé, and C. F. de Larrea, "A serum microRNA signature associated with complete remission and progression after autologous stem-cell transplantation in patients with multiple myeloma," *Oncotarget*, vol. 6, no. 3, pp. 1874–1883, 2015.

[151] J. C. Brase, M. Johannes, T. Schlomm, A. Haese, T. Steuber, T. Beissbarth, R. Kuner, and H. Sültmann, "Circulating miRNAs are correlated with tumor progression in prostate cancer," *International Journal of Cancer*, vol. 128, pp. 608–616, feb 2011.

[152] K. C. Vickers, B. T. Palmisano, B. M. Shoucri, R. D. Shamburek, and A. T. Remaley, "MicroRNAs are transported in plasma and delivered

to recipient cells by high-density lipoproteins," *Nature Cell Biology*, vol. 13, no. 4, pp. 423–435, 2011.

[153] J. D. Arroyo, J. R. Chevillet, E. M. Kroh, I. K. Ruf, C. C. Pritchard, D. F. Gibson, P. S. Mitchell, C. F. Bennett, E. L. Pogosova-Agadjanyan, D. L. Stirewalt, J. F. Tait, and M. Tewari, "Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 5003–5008, mar 2011.

[154] L. Quirico and F. Orso, "The power of microRNAs as diagnostic and prognostic biomarkers in liquid biopsies," *Cancer Drug Resistance*, vol. 3, pp. 117–139, feb 2020.

[155] M. Sochor, P. Basova, M. Pesta, N. Dusilkova, J. Bartos, P. Burda, V. Pospisil, and T. Stopka, "Oncogenic MicroRNAs: MiR-155, miR-19a, miR-181b, and miR-24 enable monitoring of early breast cancer in serum," *BMC Cancer*, vol. 14, jun 2014.

[156] C. Roth, B. Rack, V. Müller, W. Janni, K. Pantel, and H. Schwarzenbach, "Circulating microRNAs as blood-based markers for patients with primary and metastatic breast cancer," *Breast Cancer Research*, vol. 12, nov 2010.

[157] M. Zhu, Z. Huang, D. Zhu, X. Zhou, X. Shan, L. W. Qi, L. Wu, W. Cheng, J. Zhu, L. Zhang, H. Zhang, Y. Chen, W. Zhu, T. Wang, and P. Liu, "A panel of microRNA signature in serum for colorectal cancer diagnosis," *Oncotarget*, vol. 8, no. 10, pp. 17081–17091, 2017.

[158] M. L. Wikberg, R. Myte, R. Palmqvist, B. van Guelpen, and I. Ljuslinder, "Plasma miRNA can detect colorectal cancer, but how early?," *Cancer Medicine*, vol. 7, pp. 1697–1705, may 2018.

[159] Q. Wang, P. Li, A. Li, W. Jiang, H. Wang, J. Wang, and K. Xie, "Plasma specific miRNAs as predictive biomarkers for diagnosis and prognosis of glioma," *Journal of Experimental and Clinical Cancer Research*, vol. 31, no. 1, 2012.

[160] J. Zhou, L. Yu, X. Gao, J. Hu, J. Wang, Z. Dai, J. F. Wang, Z. Zhang, S. Lu, X. Huang, Z. Wang, S. Qiu, X. Wang, G. Yang, H. Sun, Z. Tang, Y. Wu, H. Zhu, and J. Fan, "Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma," *Journal of Clinical Oncology*, vol. 29, no. 36, pp. 4781–4788, 2011.

[161] D. Zheng, S. Haddadin, Y. Wang, L. Q. Gu, M. C. Perry, C. E. Freter, and M. X. Wang, "Plasma micrornas as novel biomarkers for early detection of lung cancer," *International Journal of Clinical and Experimental Pathology*, vol. 4, no. 6, pp. 575–586, 2011.

[162] G. L. Shi, Y. Chen, Y. Sun, Y. J. Yin, and C. X. Song, "Significance of serum MicroRNAs in the auxiliary diagnosis of non-small cell lung cancer," *Clinical Laboratory*, vol. 63, no. 1, pp. 133–140, 2017.

[163] Y. Zhu, J. Wang, F. Wang, Z. Yan, G. Liu, Y. Ma, W. Zhu, Y. Li, L. Xie, A. V. Bazhin, and X. Guo, "Differential MicroRNA Expression Profiles as Potential Biomarkers for Pancreatic Ductal Adenocarcinoma," *Biochemistry (Moscow)*, vol. 84, pp. 575–582, may 2019.

[164] E. Vila-Navarro, S. Duran-Sanchon, M. Vila-Casadesús, L. Moreira, A. Gins, M. Cuatrecasas, J. Josè Lozano, L. Bujanda, A. Castells, and M. Gironella, "Novel circulating mirna signatures for early detection of pancreatic neoplasia," *Clinical and Translational Gastroenterology*, vol. 10, no. 4, 2019.

[165] L. Gong, C. Wang, Y. Gao, and J. Wang, "Decreased expression of microRNA-148a predicts poor prognosis in ovarian cancer and associates with tumor growth and metastasis," *Biomedicine and Pharmacotherapy*, vol. 83, pp. 58–63, oct 2016.

[166] É. Márton, J. Lukács, A. Penyige, E. Janka, L. Hegedüs, B. Soltész, G. Méhes, R. Póka, B. Nagy, and M. Szilágyi, "Circulating epithelial-mesenchymal transition-associated miRNAs are promising biomarkers in ovarian cancer," *Journal of Biotechnology*, vol. 297, pp. 58–65, may 2019.

[167] M. S. Stark, K. Klein, B. Weide, L. E. Haydu, A. Pflugfelder, Y. H. Tang, J. M. Palmer, D. C. Whiteman, R. A. Scolyer, G. J. Mann, J. F. Thompson, G. V. Long, A. P. Barbour, H. P. Soyer, C. Garbe, A. Herington, P. M. Pollock, and N. K. Hayward, "The Prognostic and Predictive Value of Melanoma-related MicroRNAs Using Tissue and Serum: A MicroRNA Expression Analysis," *EBioMedicine*, vol. 2, pp. 671–680, jul 2015.

[168] E. Greenberg, M. J. Besser, E. Ben-Ami, R. Shapira-Frommer, O. Itzhaki, D. Zikich, D. Levy, A. Kubi, E. Eyal, A. Onn, Y. Cohen, I. Barshack, J. Schachter, and G. Markel, "A comparative analysis of

total serum miRNA profiles identifies novel signature that is highly indicative of metastatic melanoma: A pilot study," *Biomarkers*, vol. 18, pp. 502–508, sep 2013.

[169] S. Fogli, B. Polini, S. Carpi, B. Pardini, A. Naccarati, N. Dubbini, M. Lanza, M. C. Breschi, A. Romanini, and P. Nieri, "Identification of plasma microRNAs as new potential biomarkers with high diagnostic power in human cutaneous melanoma," *Tumor Biology*, vol. 39, may 2017.

[170] R. Shiiyama, S. Fukushima, M. Jinnin, J. Yamashita, A. Miyashita, S. Nakahara, A. Kogi, J. Aoi, S. Masuguchi, Y. Inoue, and H. Ihn, "Sensitive detection of melanoma metastasis using circulating microRNA expression profiles," *Melanoma Research*, vol. 23, pp. 366–372, oct 2013.

[171] C. Solé, D. Tramonti, M. Schramm, I. Goicoechea, M. Armesto, L. I. Hernandez, L. Manterola, M. Fernandez-Mercado, K. Mujika, A. Tuneu, A. Jaka, M. Tellaetxe, M. R. Friedländer, X. Estivill, P. Piazza, P. L. Ortiz-Romero, M. R. Middleton, and C. H. Lawrie, "The circulating transcriptome as a source of biomarkers for melanoma," *Cancers*, vol. 11, no. 1, 2019.

[172] C. M. Balch, J. E. Gershenwald, S. J. Soong, J. F. Thompson, M. B. Atkins, D. R. Byrd, A. C. Buzaid, A. J. Cochran, D. G. Coit, S. Ding, A. M. Eggermont, K. T. Flaherty, P. A. Gimotty, J. M. Kirkwood, K. M. McMasters, M. C. Mihm, D. L. Morton, M. I. Ross, A. J. Sober, and V. K. Sondak, "Final version of 2009 AJCC melanoma stag-

ing and classification," *Journal of Clinical Oncology*, vol. 27, no. 36, pp. 6199–6206, 2009.

[173] T. M. Morgan, "Liquid biopsy: Where did it come from, what is it, and where is it going?," *Investigative and Clinical Urology*, vol. 60, pp. 139–141, may 2019.

[174] U. Malapelle, M. Tiseo, A. Vivancos, J. Kapp, M. J. Serrano, and M. Tiemann, "Liquid Biopsy for Biomarker Testing in Non-Small Cell Lung Cancer: A European Perspective," *Journal of Molecular Pathology*, vol. 2, no. 3, pp. 255–273, 2021.

[175] M. Arechederra, M. A. Ávila, and C. Berasain, "Liquid biopsy for cancer management: a revolutionary but still limited new tool for precision medicine," *Advances in Laboratory Medicine / Avances en Medicina de Laboratorio*, vol. 1, sep 2020.

[176] Z. Zhang, S. Wu, D. L. Stenoien, and L. Paša-Tolić, "High-throughput proteomics," *Annual Review of Analytical Chemistry*, vol. 7, pp. 427–454, jul 2014.

[177] H. Ouldali, K. Sarthak, T. Ensslen, F. Piguet, P. Manivet, J. Pelta, J. C. Behrends, A. Aksimentiev, and A. Oukhaled, "Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore," *Nature Biotechnology*, vol. 38, no. 2, pp. 176–181, 2020.

[178] J. Yuan, J. Zhou, Z. Dong, S. Tandon, D. Kuk, K. S. Panageas, P. Wong, X. Wu, J. Naidoo, D. B. Page, J. D. Wolchok, and F. S. Hodi, "Pretreatment serum VEGF is associated with clinical response

and overall survival in advanced melanoma patients treated with ipilimumab," *Cancer immunology research*, vol. 2, pp. 127–132, feb 2014.

[179] T. Nedjadi, H. Benabdelkamal, N. Albarakati, A. Masood, A. Al-Sayyad, A. A. Alfadda, I. O. Alanazi, A. Al-Ammari, and J. Al-Maghrabi, "Circulating proteomic signature for detection of biomarkers in bladder cancer patients," *Scientific Reports*, vol. 10, pp. 1–10, jul 2020.

[180] J. L. Griffin and J. P. Shockcor, "Metabolic profiles of cancer cells," *Nature Reviews Cancer 2004 4:7*, vol. 4, no. 7, pp. 551–561, 2004.

[181] S. Singhal, C. Rolfo, A. W. Maksymiuk, P. S. Tappia, D. S. Sitar, A. Russo, P. S. Akhtar, N. Khatun, P. Rahnuma, A. Rashiduzzaman, R. A. Bux, G. Huang, and B. Ramjiawan, "Liquid biopsy in lung cancer screening: The contribution of metabolomics. results of a pilot study," *Cancers*, vol. 11, no. 8, 2019.

[182] B. Lee, I. Mahmud, J. Marchica, P. Dereziński, F. Qi, F. Wang, P. Joshi, F. Valerio, I. Rivera, V. Patel, C. P. Pavlovich, T. J. Garrett, G. P. Schroth, Y. Sun, and R. J. Perera, "Integrated RNA and metabolite profiling of urine liquid biopsies for prostate cancer biomarker discovery," *Scientific Reports*, vol. 10, pp. 1–17, feb 2020.

[183] D. P. Bartel, "Metazoan MicroRNAs," *Cell*, vol. 173, pp. 20–51, mar 2018.

[184] A. Grimson, M. Srivastava, B. Fahey, B. J. Woodcroft, H. R. Chiang, N. King, B. M. Degnan, D. S. Rokhsar, and D. P. Bartel, "Early

origins and evolution of microRNAs and Piwi-interacting RNAs in animals," *Nature 2008 455:7217*, vol. 455, pp. 1193–1197, oct 2008.

[185] K. U. Tüfekci, M. G. Öner, R. L. J. Meuwissen, and e. Genç, "The role of microRNAs in human diseases," *Methods in Molecular Biology*, vol. 1107, pp. 33–50, 2014.

[186] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[187] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun, "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA," *Nature 2000 408:6808*, vol. 408, pp. 86–89, nov 2000.

[188] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "MiRBase: From microRNA sequences to function," *Nucleic Acids Research*, vol. 47, no. D1, pp. D155–D162, 2019.

[189] B. Fromm, D. Domanska, E. Høye, V. Ovchinnikov, W. Kang, E. Aparicio-Puerta, M. Johansen, K. Flatmark, A. Mathelier, E. Hovig, M. Hackenberg, M. R. Friedländer, and K. J. Peterson, "MirGeneDB 2.0: The metazoan microRNA complement," *Nucleic Acids Research*, vol. 48, no. D1, pp. D132–D141, 2020.

[190] R. C. Friedman, K. K. H. Farh, C. B. Burge, and D. P. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," *Genome Research*, vol. 19, no. 1, pp. 92–105, 2009.

[191] J. P. Broughton, M. T. Lovci, J. L. Huang, G. W. Yeo, and A. E. Pasquinelli, "Pairing beyond the Seed Supports MicroRNA Targeting Specificity," *Molecular Cell*, vol. 64, pp. 320–333, oct 2016.

[192] S. Vasudevan, "Posttranscriptional Upregulation by MicroRNAs," *Wiley Interdisciplinary Reviews: RNA*, vol. 3, pp. 311–330, may 2012.

[193] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[194] D. C. Ellwanger, F. A. Büttner, H. W. Mewes, and V. Stümpflen, "The sufficient minimal set of miRNA seed types," *Bioinformatics*, vol. 27, pp. 1346–1350, may 2011.

[195] R. J. Sims, S. S. Mandal, and D. Reinberg, "Recent highlights of RNA-polymerase-II-mediated transcription," *Current opinion in cell biology*, vol. 16, pp. 263–271, jun 2004.

[196] T. A. Nguyen, M. H. Jo, Y. G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J. S. Woo, "Functional anatomy of the human microprocessor," *Cell*, vol. 161, pp. 1374–1387, jun 2015.

[197] A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello, "Genes and mechanisms related to RNA interference regulate expression of the small temporal

RNAs that control C. elegans developmental timing," *Cell*, vol. 106, no. 1, pp. 23–34, 2001.

[198] T. Kawamata and Y. Tomari, "Making RISC," *Trends in Biochemical Sciences*, vol. 35, pp. 368–376, jul 2010.

[199] D. P. Bartel, "Metazoan micrornas," *Cell*, vol. 173, no. 1, pp. 20–51, 2018.

[200] Y. Peng and C. M. Croce, "The role of microRNAs in human cancer," *Signal Transduction and Targeted Therapy*, vol. 1, no. 1, pp. 1–9, 2016.

[201] G. A. Calin, C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, and C. M. Croce, "Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 15524–15529, nov 2002.

[202] A. Cimmino, G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, R. I. Aqeilan, S. Zupo, M. Dono, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini, and C. M. Croce, "miR-15 and miR-16 induce apoptosis by targeting BCL2," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 13944–13949, sep 2005.

[203] U. Klein, M. Lia, M. Crespo, R. Siegel, Q. Shen, T. Mo, A. Ambesi-Impiombato, A. Califano, A. Migliazza, G. Bhagat, and R. Dalla-

Favera, "The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia," *Cancer cell*, vol. 17, no. 1, pp. 28–40, 2010.

[204] C. M. Croce, "Oncogenes and cancer," *The New England journal of medicine*, vol. 358, no. 5, pp. 502–511, 2008.

[205] H. Tagawa and M. Seto, "A microRNA cluster as a target of genomic amplification in malignant lymphoma," *Leukemia*, vol. 19, no. 11, pp. 2013–2016, 2005.

[206] Y. Hayashita, H. Osada, Y. Tatematsu, H. Yamada, K. Yanagisawa, S. Tomida, Y. Yatabe, K. Kawahara, Y. Sekido, and T. Takahashi, "A polycistronic MicroRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation," *Cancer Research*, vol. 65, pp. 9628–9632, nov 2005.

[207] K. J. Mavrakis, A. L. Wolfe, E. Oricchio, T. Palomero, K. De Keersmaecker, K. McJunkin, J. Zuber, T. James, K. Chang, A. A. Khan, C. S. Leslie, J. S. Parker, P. J. Paddison, W. Tam, A. Ferrando, and H. G. Wendel, "Genome-wide RNA-mediated interference screen identifies miR-19 targets in Notch-induced T-cell acute lymphoblastic leukaemia," *Nature cell biology*, vol. 12, pp. 372–379, apr 2010.

[208] C. Lu, B. C. Meyers, and P. J. Green, "Construction of small RNA cDNA libraries for deep sequencing," *Methods*, vol. 43, pp. 110–117, oct 2007.

[209] R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," *Nucleic Acids Research*, vol. 39, p. D19, jan 2011.

[210] D. C. Rio, M. Ares, G. J. Hannon, and T. W. Nilsen, "Purification of RNA using TRIzol (TRI Reagent)," *Cold Spring Harbor Protocols*, vol. 5, jun 2010.

[211] M. A. McAlexander, M. J. Phillips, and K. W. Witwer, "Comparison of methods for miRNA extraction from plasma and quantitative recovery of RNA from cerebrospinal fluid," *Frontiers in Genetics*, vol. 4, no. MAY, p. 83, 2013.

[212] V. El-Khoury, S. Pierson, T. Kaoma, F. Bernardin, and G. Berchem, "Assessing cellular and circulating miRNA recovery: The impact of the RNA isolation method and the quantity of input material," *Scientific Reports*, vol. 6, 2016.

[213] K. Wright, K. de Silva, A. C. Purdie, and K. M. Plain, "Comparison of methods for miRNA isolation and quantification from ovine plasma," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[214] C. A. Raabe, T. H. Tang, J. Brosius, and T. S. Rozhdestvensky, "Biases in small RNA deep sequencing data," *Nucleic Acids Research*, vol. 42, pp. 1414–1426, feb 2014.

[215] C. Wright, A. Rajpurohit, E. E. Burke, C. Williams, L. Collado-Torres, M. Kimos, N. J. Brandon, A. J. Cross, A. E. Jaffe, D. R. Weinberger, and J. H. Shin, "Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods," *BMC Genomics*, vol. 20, pp. 1–21, jun 2019.

[216] "Illumina Sequencing by Synthesis - YouTube."

[217] B. Ewing and P. Green, "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities," *Genome Research*, vol. 8, pp. 186–194, mar 1998.

[218] E. Aparicio-Puerta, R. Lebrón, A. Rueda, C. Gómez-Martín, S. Giannoukakos, D. Jaspez, J. M. Medina, A. Zubkovic, I. Jurak, B. Fromm, J. A. Marchal, J. Oliver, and M. Hackenberg, "SRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression," *Nucleic Acids Research*, vol. 47, no. W1, pp. W530–W535, 2019.

[219] E. Aparicio-Puerta, B. Fromm, M. Hackenberg, and M. K. Halushka, "In Silico Analysis of Micro-RNA Sequencing Data," *Methods in Molecular Biology*, vol. 2284, pp. 231–251, 2021.

[220] Y. Lu, A. S. Baras, and M. K. Halushka, "miRge 2.0 for comprehensive analysis of microRNA sequencing data," *BMC bioinformatics*, vol. 19, jul 2018.

[221] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in Drosophila.," *Genome biology*, vol. 5, no. 1, 2003.

[222] F. Kern, E. Aparicio-Puerta, Y. Li, T. Fehlmann, T. Kehl, V. Wagner, K. Ray, N. Ludwig, H. P. Lenhof, E. Meese, and A. Keller, "miRTargetLink 2.0 - Interactive miRNA target gene and target pathway networks," *Nucleic Acids Research*, vol. 49, pp. W409–W416, jul 2021.

[223] T. R. Cech and J. A. Steitz, "The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones," *Cell*, vol. 157, pp. 77–94, mar 2014.

[224] R. D. Morin, M. D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A. L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C. J. Eaves, and M. A. Marra, "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, pp. 610–621, apr 2008.

[225] N. Cloonan, S. Wani, Q. Xu, J. Gu, K. Lea, S. Heater, C. Barbacioru, A. L. Steptoe, H. C. Martin, E. Nourbakhsh, K. Krishnan, B. Gardiner, X. Wang, K. Nones, J. A. Steen, N. A. Matigian, D. L. Wood, K. S. Kassahn, N. Waddell, J. Shepherd, C. Lee, J. Ichikawa, K. McKernan, K. Bramlett, S. Kuersten, and S. M. Grimmond, "MicroRNAs and their isomiRs function cooperatively to target common biological pathways," *Genome Biology*, vol. 12, no. 12, pp. 1–20, 2011.

[226] C. W. Wu, J. M. Evans, S. Huang, D. W. Mahoney, B. A. Dukek, W. R. Taylor, T. C. Yab, T. C. Smyrk, J. Jen, J. B. Kisiel, and D. A. Ahlquist, "A Comprehensive Approach to Sequence-oriented IsomiR annotation (CASMIR): Demonstration with IsomiR profiling in colorectal neoplasia," *BMC Genomics*, vol. 19, pp. 1–12, may 2018.

[227] G. C. Tan, E. Chan, A. Molnar, R. Sarkar, D. Alexieva, I. M. Isa, S. Robinson, S. Zhang, P. Ellis, C. F. Langford, P. V. Guillot, A. Chandrashekran, N. M. Fisk, L. Castellano, G. Meister, R. M. Winston, W. Cui, D. Baulcombe, and N. J. Dibb, "5' isomiR vari-

ation is of functional and evolutionary importance," *Nucleic Acids Research*, vol. 42, no. 14, pp. 9424–9435, 2014.

[228] A. G. Telonis, R. Magee, P. Loher, I. Chervoneva, E. Londin, and I. Rigoutsos, "Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types," *Nucleic Acids Research*, vol. 45, no. 6, pp. 2973–2985, 2017.

[229] X. Bofill-De Ros, A. Yang, and S. Gu, "IsomiRs: Expanding the miRNA repression toolbox beyond the seed," *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1863, no. 4, 2020.

[230] A. Yang, X. Bofill-De Ros, T. J. Shao, M. Jiang, K. Li, P. Villanueva, L. Dai, and S. Gu, "3' Uridylation Confers miRNAs with Non-canonical Target Repertoires," *Molecular Cell*, vol. 75, pp. 511–522.e4, aug 2019.

[231] E. Aparicio-Puerta, C. Gomez-Martin, S. Giannoukakos, J. Maria Medina, J. A. Marchal, and M. Hackenberg, "mirnaQC: A webserver for comparative quality control of miRNA-seq data," *Nucleic Acids Research*, vol. 48, pp. W262–W267, jul 2020.

[232] "omicidx/omicidx-api."

[233] "SRA Tools."

[234] Student, "The Probable Error of a Mean," *Biometrika*, vol. 6, p. 1, mar 1908.

[235] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, p. 139, nov 2010.

[236] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, pp. 1–9, mar 2010.

[237] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, pp. 1–12, oct 2010.

[238] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, pp. 1–21, 2014.

[239] F. Seyednasrollah, A. Laiho, and L. L. Elo, "Comparison of software packages for detecting differential expression in RNA-seq studies," *Briefings in Bioinformatics*, vol. 16, no. 1, pp. 59–70, 2013.

[240] S. Tarazona, P. Furió-Tarí, D. Turrà, A. Di Pietro, M. J. Nueda, A. Ferrer, and A. Conesa, "Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package," *Nucleic Acids Research*, vol. 43, no. 21, 2015.

[241] H. Lan, H. Lu, X. Wang, and H. Jin, "MicroRNAs as potential biomarkers in cancer: Opportunities and challenges," *BioMed Research International*, vol. 2015, 2015.

[242] J. Hayes, P. P. Peruzzi, and S. Lawler, "MicroRNAs in cancer: Biomarkers, functions and therapy," *Trends in Molecular Medicine*, vol. 20, no. 8, pp. 460–469, 2014.

[243] 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 Research*, vol. 40, no. 1, pp. 37–52, 2012.

[244] M. J. Axtell, "ShortStack: Comprehensive annotation and quantification of small RNA genes," *Rna*, vol. 19, no. 6, pp. 740–751, 2013.

[245] L. Pantano, X. Estivill, and E. Martí, "SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells," *Nucleic Acids Research*, vol. 38, no. 5, 2009.

[246] G. Barturen, A. Rueda, M. Hamberg, A. Alganza, R. Lebron, M. Kotsyfakis, B.-J. Shi, D. Koppers-Lalic, and M. Hackenberg, "sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments," *Methods in Next Generation Sequencing*, vol. 1, no. 1, pp. 21–31, 2014.

[247] W. Kang, Y. Eldfjell, B. Fromm, X. Estivill, I. Biryukova, and M. R. Friedländer, "MiRTrace reveals the organismal origins of microRNA sequencing data," *Genome Biology*, vol. 19, no. 1, pp. 1–15, 2018.

[248] A. Rueda, G. Barturen, R. Lebrón, C. Gómez-Martín, Á. Alganza, J. L. Oliver, and M. Hackenberg, "SRNAtoolbox: An integrated col-

lection of small RNA research tools," *Nucleic Acids Research*, vol. 43, no. W1, pp. W467–W473, 2015.

[249] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 178–192, 2013.

[250] T. Galili, A. O'Callaghan, J. Sidi, and C. Sievert, "Heatmaply: An R package for creating interactive cluster heatmaps for online publishing," *Bioinformatics*, vol. 34, no. 9, pp. 1600–1602, 2018.

[251] J. R. Conway, A. Lex, and N. Gehlenborg, "UpSetR: An R package for the visualization of intersecting sets and their properties," *Bioinformatics*, vol. 33, pp. 2938–2940, sep 2017.

[252] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek, "Ensembl 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D754–D761, 2018.

[253] R. Agarwala, T. Barrett, J. Beck, D. A. Benson, C. Bollin, E. Bolton, D. Bourexis, J. R. Brister, S. H. Bryant, K. Canese, C. Charowhas, K. Clark, M. Dicuccio, I. Dondoshansky, S. Federhen, M. Feolo, K. Funk, L. Y. Geer, V. Gorelenkov, M. Hoeppner, B. Holmes, M. Johnson, V. Khotomlianski, A. Kimchi, M. Kimelman, P. Kitts, W. Klimke, S. Krasnov, A. Kuznetsov, M. J. Landrum, D. Landsman, J. M. Lee, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, A. Marchler-Bauer, I. Karsch-Mizrachi, T. Murphy, R. Orris, J. Ostell, C. O'sullivan, A. Panchenko, L. Phan, D. Preuss, K. D. Pruitt, K. Rodarmer, W. Rubinstein, E. Sayers, V. Schneider, G. D. Schuler, S. T. Sherry, K. Sirotkin, K. Siyan, D. Slotta, A. Soboleva, V. Soussov, G. Starchenko, T. A. Tatusova, K. Todorov, B. W. Trawick, D. Vakatov, Y. Wang, M. Ward, W. J. Wilbur, E. Yaschenko, and K. Zbicz, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 44, no. D1, pp. D7–D19, 2016.

[254] B. Fromm, T. Billipp, L. E. Peck, M. Johansen, J. E. Tarver, B. L. King, J. M. Newcomb, L. F. Sempere, K. Flatmark, E. Hovig, and K. J. Peterson, "A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome," *Annual Review of Genetics*, vol. 49, no. 1, pp. 213–242, 2015.

[255] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, 2009.

[256] M. Hackenberg, N. Rodríguez-Ezpeleta, and A. M. Aransay, "MiR-

analyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments," *Nucleic Acids Research*, vol. 39, no. SUPPL. 2, pp. 132–138, 2011.

[257] M. J. Axtell, "Classification and comparison of small RNAs from plants," *Annual Review of Plant Biology*, vol. 64, pp. 137–159, 2013.

[258] R. M. Kuhn, D. Haussler, and W. James Kent, "The UCSC genome browser and associated tools," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 144–161, 2013.

[259] D. Koppers-Lalic, M. Hackenberg, I. V. Bijnsdorp, M. A. van Eijndhoven, P. Sadek, D. Sie, N. Zini, J. M. Middeldorp, B. Ylstra, R. X. de Menezes, T. Würdinger, G. A. Meijer, and D. M. Pegtel, "Nontemplated nucleotide additions distinguish the small RNA composition in cells from exosomes," *Cell Reports*, vol. 8, pp. 1649–1658, sep 2014.

[260] K. Pachmann, O. Camara, T. Kroll, M. Gajda, A. K. Gellner, J. Wotschadlo, and I. B. Runnebaum, "Efficacy control of therapy using circulating epithelial tumor cells (CETC) as "liquid biopsy": trastuzumab in HER2/neu-positive breast carcinoma," *Journal of cancer research and clinical oncology*, vol. 137, pp. 1317–1327, sep 2011.

[261] E. Heitzer, S. Perakis, J. B. Geigl, and M. R. Speicher, "The potential of liquid biopsies for the early detection of cancer," *NPJ precision oncology*, vol. 1, dec 2017.

[262] A. Esquela-Kerscher and F. J. Slack, "Oncomirs - microRNAs with a role in cancer," *Nature reviews. Cancer*, vol. 6, pp. 259–269, apr 2006.

[263] A. Michael, S. D. Bajracharya, P. S. Yuen, H. Zhou, R. A. Star, G. G. Illei, and I. Alevizos, "Exosomes from human saliva as a source of microRNA biomarkers," *Oral diseases*, vol. 16, pp. 34–38, jan 2010.

[264] T. Yuan, X. Huang, M. Woodcock, M. Du, R. Dittmar, Y. Wang, S. Tsai, M. Kohli, L. Boardman, T. Patel, and L. Wang, "Plasma extracellular RNA profiles in healthy and cancer patients," *Scientific reports*, vol. 6, jan 2016.

[265] X. Huang, T. Yuan, M. Tschannen, Z. Sun, H. Jacob, M. Du, M. Liang, R. L. Dittmar, Y. Liu, M. Liang, M. Kohli, S. N. Thibodeau, L. Boardman, and L. Wang, "Characterization of human plasma-derived exosomal RNAs by deep sequencing," *BMC genomics*, vol. 14, may 2013.

[266] S. Uhlmann, E. Mracsko, E. Javidi, S. Lamble, A. Teixeira, A. Hotz-Wagenblatt, K. H. Glatting, and R. Veltkamp, "Genome-Wide Analysis of the Circulating miRNome After Cerebral Ischemia Reveals a Reperfusion-Induced MicroRNA Cluster," *Stroke*, vol. 48, pp. 762–769, mar 2017.

[267] M. Seco-Cervera, D. González-Rodríguez, J. S. Ibáñez-Cabellos, L. Peiró-Chova, F. V. Pallardó, and J. L. García-Giménez, "Small RNA-seq analysis of circulating miRNAs to identify phenotypic vari-

ability in Friedreich's ataxia patients," *Scientific data*, vol. 5, mar 2018.

[268] C. N. Correia, N. C. Nalpas, K. E. McLoughlin, J. A. Browne, S. V. Gordon, D. E. MacHugh, and R. G. Shaughnessy, "Circulating microRNAs as Potential Biomarkers of Infectious Disease," *Frontiers in immunology*, vol. 8, feb 2017.

[269] E. M. Kroh, R. K. Parkin, P. S. Mitchell, and M. Tewari, "Analysis of circulating microRNA biomarkers in plasma and serum using quantitative reverse transcription-PCR (qRT-PCR)," *Methods (San Diego, Calif.)*, vol. 50, pp. 298–301, apr 2010.

[270] F. Russo, S. Di Bella, F. Vannini, G. Berti, F. Scoyni, H. V. Cook, A. Santos, G. Nigita, V. Bonnici, A. Laganà, F. Geraci, A. Pulvirenti, R. Giugno, F. De Masi, K. Belling, L. J. Jensen, S. Brunak, M. Pellegrini, and A. Ferro, "miRandola 2017: a curated knowledge base of non-invasive biomarkers," *Nucleic acids research*, vol. 46, pp. D354–D359, jan 2018.

[271] S. Mathivanan and R. J. Simpson, "ExoCarta: A compendium of exosomal proteins and RNA," *Proteomics*, vol. 9, pp. 4997–5000, nov 2009.

[272] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic acids research*, vol. 42, jan 2014.

[273] P. P. Chan and T. M. Lowe, "GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes," *Nucleic acids research*, vol. 44, no. D1, pp. D184–D189, 2016.

[274] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 35, jan 2007.

[275] A. I. Petrov, S. J. Kay, I. Kalvari, K. L. Howe, K. A. Gray, E. A. Bruford, P. J. Kersey, G. Cochrane, R. D. Finn, A. Bateman, A. Kozomara, S. Griffiths-Jones, A. Frankish, C. W. Zwieb, B. Y. Lau, K. P. Williams, P. P. Chan, T. M. Lowe, J. J. Cannone, R. R. Gutell, M. A. Machnicka, J. M. Bujnicki, M. Yoshihama, N. Kenmochi, B. Chai, J. R. Cole, M. Szymanski, W. M. Karlowski, V. Wood, E. Huala, T. Z. Berardini, Y. Zhao, R. Chen, W. Zhu, M. D. Paraskevopoulou, I. S. Vlachos, A. G. Hatzigeorgiou, L. Ma, Z. Zhang, J. Puetz, P. F. Stadler, D. McDonald, S. Basu, P. Fey, S. R. Engel, J. M. Cherry, P. J. Volders, P. Mestdagh, J. Wower, M. Clark, X. C. Quek, and M. E. Dinger, "RNAcentral: a comprehensive database of non-coding RNA sequences," *Nucleic acids research*, vol. 45, pp. D128–D134, jan 2017.

[276] A. Baraniskin, S. Nöpel-Dünnebacke, M. Ahrens, S. G. Jensen, H. Zöllner, A. Maghnouj, A. Wos, J. Mayerle, J. Munding, D. Kost, A. Reinacher-Schick, S. Liffers, R. Schroers, A. M. Chromik, H. E. Meyer, W. Uhl, S. Klein-Scory, F. U. Weiss, C. Stephan, I. Schwarte-Waldhoff, M. M. Lerch, A. Tannapfel, W. Schmiegel, C. L. Andersen, and S. A. Hahn, "Circulating U2 small nuclear RNA fragments as a

novel diagnostic biomarker for pancreatic and colorectal adenocarcinoma," *International journal of cancer*, vol. 132, jan 2013.

[277] M. D. Giraldez, R. M. Spengler, A. Etheridge, P. M. Godoy, A. J. Barczak, S. Srinivasan, P. L. De Hoff, K. Tanriverdi, A. Courtright, S. Lu, J. Khoory, R. Rubio, D. Baxter, T. A. Driedonks, H. P. Buermans, E. N. Nolte-'T Hoen, H. Jiang, K. Wang, I. Ghiran, Y. E. Wang, K. Van Keuren-Jensen, J. E. Freedman, P. G. Woodruff, L. C. Laurent, D. J. Erle, D. J. Galas, and M. Tewari, "Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling," *Nature biotechnology*, vol. 36, pp. 746–757, sep 2018.

[278] V. Balatti, Y. Pekarsky, and C. M. Croce, "Role of the tRNA-Derived Small RNAs in Cancer: New Potential Biomarkers and Target for Therapy," *Advances in cancer research*, vol. 135, pp. 173–187, 2017.

[279] D. Koppers-Lalic, M. Hackenberg, R. De Menezes, B. Misovic, M. Wachalska, A. Geldof, N. Zini, T. De Reijke, T. Wurdinger, A. Vis, J. Van Moorselaar, M. Pegtel, and I. Bijnsdorp, "Non-invasive prostate cancer detection by measuring miRNA variants (isomiRs) in urine extracellular vesicles," *Oncotarget*, vol. 7, pp. 22566–22578, apr 2016.

[280] K. Sorefan, H. Pais, A. E. Hall, A. Kozomara, S. Griffiths-Jones, V. Moulton, and T. Dalmay, "Reducing ligation bias of small RNAs in libraries for next generation sequencing," *Silence*, vol. 3, may 2012.

[281] P. Tiberio, M. Callari, V. Angeloni, M. G. Daidone, and V. Appierto, "Challenges in Using Circulating miRNAs as Cancer Biomarkers," *BioMed Research International*, vol. 2015, p. 731479, 2015.

[282] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, "Normalization of RNA-seq data using factor analysis of control genes or samples," *Nature biotechnology*, vol. 32, pp. 896–902, sep 2014.

[283] S. Shore, J. M. Henderson, A. Lebedev, M. P. Salcedo, G. Zon, A. P. McCaffrey, N. Paul, and R. I. Hogrefe, "Small RNA Library Preparation Method for Next-Generation Sequencing Using Chemical Modifications to Prevent Adapter Dimer Formation," *PloS one*, vol. 11, nov 2016.

[284] A. M. Plocik and B. R. Graveley, "New insights from existing sequence data: generating breakthroughs without a pipette," *Molecular cell*, vol. 49, pp. 605–617, feb 2013.

[285] "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data."

[286] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics (Oxford, England)*, vol. 32, pp. 3047–3048, oct 2016.

[287] D. Witten, R. Tibshirani, S. G. Gu, A. Fire, and W. O. Lui, "Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls," *BMC biology*, vol. 8, may 2010.

[288] P. Łaniewski, D. Barnes, A. Goulder, H. Cui, D. J. Roe, D. M. Chase, and M. M. Herbst-Kralovetz, "Linking cervicovaginal immune signatures, HPV and microbiota composition in cervical carcinogenesis in non-Hispanic and Hispanic women," *Scientific Reports 2018 8:1*, vol. 8, pp. 1–13, may 2018.

[289] C. Wongwarangkana, K. E. Fujimori, M. Akiba, S. Kinoshita, M. Teruya, M. Nezuo, T. Masatoshi, S. Watabe, and S. Asakawa, "Deep sequencing, profiling and detailed annotation of microRNAs in Takifugu rubripes," *BMC Genomics*, vol. 16, pp. 1–14, jun 2015.

[290] American Cancer Society, "What is melanoma skin cancer?," *American Cancer Society*, pp. 1–5, 2019.

[291] M. J. Sneyd and B. Cox, "A comparison of trends in melanoma mortality in New Zealand and Australia: The two countries with the highest melanoma incidence and mortality in the world," *BMC Cancer*, vol. 13, p. 372, aug 2013.

[292] European Environment and Health Information System, "Incidence of Melanoma in People Aged Under 55 Years," vol. 2009, no. December, 2009.

[293] World Health Organization, "Cancer Factsheet Australia," 2020.

[294] American Cancer Society, "Survival Rates for Melanoma Skin Cancer," 2020.

[295] L. Kuryk, L. Bertinato, M. Staniszewska, K. Pancer, M. Wieczorek, S. Salmaso, P. Caliceti, and M. Garofalo, "From conventional therapies to immunotherapy: Melanoma treatment in review," oct 2020.

[296] D. J. Wong and A. Ribas, "Targeted therapy for melanoma," in *Cancer Treatment and Research*, vol. 167, pp. 251–262, Cancer Treat Res, 2016.

[297] E. Alegre, M. Sammamed, S. Fernández-Landázuri, L. Zubiri, and Á. González, "Circulating Biomarkers in Malignant Melanoma," in *Advances in Clinical Chemistry*, vol. 69, pp. 47–89, Elsevier, jan 2015.

[298] A. K. Chan, R. W. Chiu, and Y. M. Lo, "Cell-free nucleic acids in plasma, serum and urine: A new tool in molecular diagnosis," mar 2003.

[299] P. Leidinger, A. Keller, A. Borries, J. Reichrath, K. Rass, S. U. Jager, H. P. Lenhof, and E. Meese, "High-throughput miRNA profiling of human melanoma blood samples," *BMC Cancer*, vol. 10, pp. 1–11, jun 2010.

[300] J. L. Palacios-Ferrer, M. B. García-Ortega, M. Gallardo-Gómez, M. Á. García, C. Díaz, H. Boulaiz, J. Valdivia, J. M. Jurado, F. M. Almazan-Fernandez, S. Arias-Santiago, V. Amezcua, H. Peinado, F. Vicente, J. Pérez del Palacio, and J. A. Marchal, "Metabolomic profile of cancer stem cell-derived exosomes from patients with malignant melanoma," *Molecular Oncology*, vol. 15, pp. 407–428, feb 2021.

[301] N. Wan, W. Yang, H. Cheng, and J. Wang, "Foxd3-as1 contributes to the progression of melanoma via mir-127-3p/fjx1 axis," *https://home.liebertpub.com/cbr*, vol. 35, pp. 596–604, 10 2020.

[302] P. Tian, L. Tao, Y. Wang, and X. Han, "Microrna-127 inhibits the progression of melanoma by downregulating delta-like homologue 1," *BioMed Research International*, vol. 2020, 2020.

[303] J. Fellenberg, P. Kunz, B. Lehner, H. Saehr, and A. Schenker, "Tumor suppressor function of mir-127-3p and mir-376a-3p in osteosarcoma cells," *Cancers 2019, Vol. 11, Page 2019*, vol. 11, p. 2019, 12 2019.

[304] S. Wang, H. Li, J. Wang, D. Wang, A. Yao, and Q. Li, "Prognostic and biological significance of microrna-127 expression in human breast cancer," *Disease Markers*, vol. 2014, 2014.

[305] S. Wei and W. Ma, "Mir-370 functions as oncogene in melanoma by direct targeting pyruvate dehydrogenase b," *Biomedicine & Pharmacotherapy*, vol. 90, pp. 278–286, 6 2017.

[306] F. Kern, T. Fehlmann, J. Solomon, L. Schwed, N. Grammes, C. Backes, K. van Keuren-Jensen, D. W. Craig, E. Meese, and A. Keller, "mieaa 2.0: integrating multi-species microrna enrichment analysis and workflow management systems," *Nucleic acids research*, vol. 48, pp. W521–W528, 2020.

[307] Z. Ding, S. Jian, X. Peng, Y. Liu, J. Wang, L. Zheng, C. Ou, Y. Wang, W. Zeng, and M. Zhou, "Loss of mir-664 expression enhances cutaneous malignant melanoma proliferation by upregulating plp2," *Medicine (United States)*, vol. 94, p. e1327, 8 2015.

[308] M. O. Urbanek-Trzeciak, P. Galka-Marciniak, P. M. Nawrocka, E. Kowal, S. Szwec, M. Giefing, and P. Kozlowski, "Pan-cancer analysis of somatic mutations in mirna genes," *EBioMedicine*, vol. 61, p. 103051, 11 2020.

[309] H. Kim, J. Kim, S. Yu, Y. Y. Lee, J. Park, R. J. Choi, S. J. Yoon, S. G. Kang, and V. N. Kim, "A mechanism for microrna arm switching regulated by uridylation," *Molecular Cell*, vol. 78, pp. 1224–1236.e5, 6 2020.

[310] B. Kim, M. Ha, L. Loeff, H. Chang, D. K. Simanshu, S. Li, M. Fareh, D. J. Patel, C. Joo, and V. N. Kim, "Tut7 controls the fate of precursor micrornas by using three different uridylation mechanisms," *The EMBO Journal*, vol. 34, p. 1801, 7 2015.

[311] L. Bi, Q. Yang, J. Yuan, Q. Miao, L. Duan, F. Li, and S. Wang, "Microrna-127-3p acts as a tumor suppressor in epithelial ovarian cancer by regulating the bag5 gene," *Oncology reports*, vol. 36, pp. 2563–2570, 11 2016.

[312] H. Jiang, C. Jin, J. Liu, D. Hua, F. Zhou, X. Lou, N. Zhao, Q. Lan, Q. Huang, J. G. Yoon, S. Zheng, and B. Lin, "Next generation sequencing analysis of mirnas: Mir-127-3p inhibits glioblastoma proliferation and activates tgf-b signaling by targeting ski," *OMICS : a Journal of Integrative Biology*, vol. 18, p. 196, 3 2014.

[313] L. Wang, X. Wang, and X. Jiang, "Mir-127 suppresses gastric cancer cell migration and invasion via targeting wnt7a," *Oncology Letters*, vol. 17, pp. 3219–3226, 3 2019.

[314] A. Russo, R. Caltabiano, A. Longo, T. Avitabile, L. M. Franco, V. Bonfiglio, L. Puzzo, and M. Reibaldi, "Increased levels of mirna-146a in serum and histologic samples of patients with uveal melanoma," *Frontiers in Pharmacology*, vol. 7, p. 424, 11 2016.

[315] M. L. Nueda, A. I. Naranjo, V. Baladrón, and J. Laborda, "The proteins dlk1 and dlk2 modulate notch1-dependent proliferation and oncogenic potential of human sk-mel-2 melanoma cells," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1843, pp. 2674–2684, 11 2014.

[316] Y. Guo, X. Zhang, L. Wang, M. Li, M. Shen, Z. Zhou, S. Zhu, K. Li, Z. Fang, B. Yan, S. Zhao, J. Su, X. Chen, and C. Peng, "The plasma exosomal mir-1180-3p serves as a novel potential diagnostic marker for cutaneous melanoma," *Cancer Cell International*, vol. 21, pp. 1–15, 12 2021.

[317] E. E. Drees, M. G. Roemer, N. J. Groenewegen, J. Perez-Boza, M. A. van Eijndhoven, L. I. Prins, S. A. Verkuijlen, X. M. Tran, J. Driessen, G. J. Zwezerijnen, P. Stathi, K. Mol, J. J. Karregat, A. Kalantidou, A. Vallés-Martí, T. J. Molenaar, E. Aparicio-Puerta, E. van Dijk, B. Ylstra, C. G. Groothuis-Oudshoorn, M. Hackenberg, D. de Jong, J. M. Zijlstra, and D. M. Pegtel, "Extracellular vesicle mirna predict fdg-pet status in patients with classical hodgkin lymphoma," *Journal of Extracellular Vesicles*, vol. 10, p. e12121, 7 2021.