Computational analysis of microRNAs in biomedicine

Katherine Abigail Icay-Rouhiainen

Research Programs Unit,
Genome-Scale Biology
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 2, Haartmaninkatu 8, Helsinki,
on 7th September 2018, at 12 o'clock noon.

Helsinki 2018

Supervisor

Sampsa Hautaniemi, DTech, Professor
Research Programs Unit, Genome-Scale Biology,
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Pre-examiners appointed by the Faculty

Jaana Hartikainen, PhD, Adjunct Professor
Faculty of Health Sciences
School of Medicine
Institute of Clinical Medicine,
Clinical Pathology and Forensic Medicine
University of Eastern Finland, Kuopio, Finland

Samuel Kaski, Professor
Helsinki Institute for Information Technology HIIT,
Department of Computer Science, Aalto University
Helsinki, Finland

Opponent appointed by the Faculty

Rolf Skotheim, Professor/Group Leader
Genome Biology Group, Department of Molecular Oncology
Institute for Cancer Research, Oslo University Hospital
Institute of Informatics, University of Oslo
Oslo, Norway

ISBN 978-951-51-4433-1 (paperback)
ISBN 978-951-51-4434-8 (PDF)
`http://ethesis.helsinki.fi`
Unigrafia
Helsinki 2018

For Papa, Mama, Sami and BB.

All scientific knowledge that we have of this world, or will ever have, is as an island in the sea of mystery. We live in our partial knowledge as the Dutch live on polders claimed from the sea. We dike and fill. We dredge up soil from the bed of mystery and build ourselves room to grow.

Chet Raymo, 1998

Abstract

All genetic information necessary for creating and maintaining life is stored in DNA and RNA molecules. Gene expression is the process by which sets of DNA (i.e. genes) are encoded into functional gene products. Thus, the state and function of a single cell can be determined by the amount and type of genes expressed: tumour cells can be detected from normal cells, and one functional brain region can be differentiated from another.

The discovery of non-coding RNAs like microRNAs (miRNAs) introduced a sophisticated level of gene regulation to our understanding of the flow of genetic information. Strong evidence suggest miRNAs have vital roles in mediating a wide range of biological pathways essential to cell maintenance and tissue-specific function. In complex diseases such as cancer, they show particular promise as candidate biomarkers in prognosis, diagnosis, and treatment. However, we are still uncertain about the precise mechanisms and contributions of miRNAs in regulating gene expression.

High-throughput technologies generate molecular data of unprecedented size and depth, providing unique opportunities to study small RNA molecules and complex diseases. Despite exact regulatory mechanisms being uncertain, miRNAs are functionally characterized with high-throughput expression data and the biological pathways annotated to their putative target genes. However, the sheer size of the data generated and to be processed raises challenges in computational resources and in discovering clinically relevant information.

This work addresses these challenges with the development and application of two computational tools to better facilitate miRNA research. SePIA is a high-throughput workflow to reliably process sequencing data and perform expression analysis to identify strongly-related miRNAs and their predicted target genes. Director is a visualization package to further the interpretation of molecular interactions and depict the co-regulatory behaviour of miRNAs.

The usefulness of these tools is shown in the application of two biomedical studies: in differentiating brain tissue phenotypes, and in determining a role in the chemosensitivity of diffuse large B-cell lymphoma. Sufficient biological context is drawn from the computational results generated by the tools to hypothesize and experimentally validate the role of miRNAs, and propose a set as candidate biomarkers and targets for drug therapy.

SePIA and Director are readily available tools developed to improve and make more convenient the computational analysis of miRNAs in biomedical research.

Contents

Abbreviations

Pub I	original publication I
Pub II	original publication II
Pub III	original publication III
Pub IV	original publication IV
cDNA	complementary DNA
COL11A1	Collagen type XI alpha I
DCN	Decorin
DE	differentially expressed
DLBCL	Diffuse large B-cell lymphoma
DNA	deoxyribonucleic acid
EEF1A2	Eukaryotic translation elongation factor 1 alpha 2
HOXA3	Homeobox A3
HOXD10	Homeobox D10
INHBA	Inhibin beta A
KLF6	Krupper-like factor 6
KRAS	Ki-ras
MAPK	Mitogen-activated protein kinase
MFE	minimum free energy
miRNA	microRNA
mRE	microRNA response element
mRNA	messenger RNA
NGS	next-generation sequencing technology
NRAS	Neuroblastoma RAS viral oncogene homolog
OVCA	High-grade serous ovarian cancer
PTEN	Phosphatase and tensin homolog
qRT-PCR	quantitative reverse transcription polymerase chain reaction
RISC	RNA-induced silencing complex
RMA	robust multi-array average
RNA	ribonucleic acid
RNA-seq	next-generation RNA sequencing technology
RNAi	RNA interference
RPKM	reads per kilobase per million mapped reads
SePIA	RNA and small-RNA sequence processing, integration, and analysis
SNORA	Small nucleolar RNA, H/ACA Box
SNORD	Small nucleolar RNA, C/D Box
TCGA	The Cancer Genome Atlas project
TGFβ	Transforming growth factor beta
TGFBR2	Transforming growth factor beta receptor 2
THBS2	Thrombospondin 2
TP53	Tumour protein 53
UTR	untranslated region

List of original publications

Publication I **Katherine Icay***, Ping Chen*, Alejandra Cervera*, Ville Rantanen, Rainer Lehtonen and Sampsa Hautaniemi.
(2016) SePIA: RNA and small RNA sequence processing, integration, and analysis. *BioData Mining* **9**, 20.

Publication II **Katherine Icay**, Chengyu Liu and Sampsa Hautaniemi.
(2018) Dynamic visualization of multi-level molecular data: the Director package in R. *Computer Methods and Programs in Biomedicine* **153**, 129–136.

Publication III Suvi-Katri Leivonen*, **Katherine Icay***, Kirsi Jäntti, Ilari Siren, Chengyu Liu, Amjad Alkodsi, Alejandra Cervera, Maja Ludvigsen, Stephen Jacques Hamilton-Dutoit, Francesco d'Amore, Marja-Liisa Karjalainen-Lindsberg, Jan Delabie, Harald Holte, Rainer Lehtonen, Sampsa Hautaniemi, Sirpa Leppä.
(2017) MicroRNAs regulate key cell survival pathways and mediate chemosensitivity during progression of diffuse large B-cell lymphoma. *Blood Cancer Journal* **7**, 654.

Publication IV Juuso Juhila, Tessa Sipilä, **Katherine Icay**, Daniel Nicorici, Pekka Ellonen, Aleksi Kallio, Eija Korpelainen, Dario Greco, Iiris Hovatta
(2011) MicroRNA expression profiling reveals miRNA families regulating specific biological pathways in mouse frontal cortex and hippocampus. *PLoS ONE* **6**, e21495.

* equal contribution to work

Publications included in other thesis

Publication III was included in the thesis of Chengyu Liu (Computational integrative analysis of biological networks in cancer, Helsinki 2017).

Author's contributions

Publication I Designed the overall workflow for processing and analysis of small RNA and integration to RNA-seq data. Developed components for the workflow, primarily for small RNA processing and expression analysis. Carried out analysis for the manuscript and executed test cases. Wrote the manuscript.

Publication II Designed and developed the R package. Collected TCGA data; performed expression correlation analysis, target prediction, and visualization for both test cases. Wrote the manuscript.

Publication III Implemented computational workflow to characterize miRNAs in matched, primary and relapsed DLBCL tumour samples. Identified functionally interesting target genes of differentially expressed miRNAs using SePIA for experimental validation. Created supplementary figures with Director and wrote the manuscript.

Publication IV Performed expression analysis of microarray and small RNA-seq data. Identified microRNAs differentially expressed between the tissue samples using microarray and small RNA-seq expression. Putative target genes were identified for selected miRNA families and clusters and used for pathway analysis. Wrote the corresponding sections of the manuscript.

1 Introduction

DNA is transcribed into RNA and RNA is translated into protein. This is the central dogma originally proposed by Francis Crick [1] to explain the flow of genetic information responsible for all known life on Earth. Proteins were the functional end product, DNA the blueprint for creating and maintaining life, and RNA the humble messenger between the two.

Some scientists, however, recognized that RNA was more than a transitional step in the flow of genetic information and noted a unique combination of features that allowed it to do things neither DNA nor protein can. The RNA World is the concept that RNAs — or something chemically similar — was the primary living substance on Earth around 4 billion years ago and 'carried out most of the information processing and metabolic transformation needed for *biology* to emerge from *chemistry*.' [2] In other words, RNA is linked to the origin of life [3]. Strong evidence also indicate RNA preceded DNA and protein: deoxyribose synthesis is from ribose and proteins lack the potential for complementary pairing that make heredity and evolution possible. Viral RNA was used to demonstrate how RNA could carry heritable information, catalyze basic reactions, self-replicate, and evolve when subject to natural selection without the presence of cells [4].

Unfortunately, the protein-centered bias inherent in the traditional central dogma meant that the discovery of large regions of DNA that transcribed to RNA but did not translate into proteins were initially dismissed as 'junk' resulting from evolutionary redundancy [5]. New understanding brought on with the advent of high-throughput technology forced a revision of the central dogma to include numerous classes of non-protein coding RNAs with functional roles in translation (e.g. ribosomalRNA. transferRNA), gene splicing (e.g. small nuclear RNA), and epigenetics (e.g. long non-coding RNAs) [6]. The introduction of these classes revealed a much more sophisticated and evolved system of molecular regulation than initially anticipated. The Encyclopedia of DNA Elements (ENCODE) project found that at least 76% of the human genome is transcribed into RNAs, with protein-coding genes making up less than 3% of the human genome [7]. Among these classes of non-protein coding RNAs are a set of small RNAs called microRNAs (miRNAs) with unprecedented functional importance in development and disease.

MiRNAs have the ability to inhibit the translation of RNA into proteins. The modest strand length averaging 22 nucleotides belies a highly-complex ability to orchestrate the regulation of hundreds of protein-coding RNA simultaneously. The exact mechanisms by which such small RNA molecules achieve so much is a mystery scientists are still unravelling, but what is known supports the concept of

RNA as the molecular foundation on which cellular and complex, multi-cellular life became possible [3].

This thesis work contributes to the active area of miRNA research elucidating biological function from expression data, with emphasis on complex and difficult-to-treat disease. In cancer, for example, strong evidence suggests some miRNAs can directly affect tumor growth and development through associated target genes [6]. A computational workflow was developed to enhance the process and analysis of high-throughput miRNA data with complementing messenger RNA (mRNA) data. MiRNA expression in two very different datasets (mouse brain tissue and human tumours) were comprehensively profiled and analyzed using existing tools implemented within a workflow to identify potential biomarkers for disease. A visualization approach was developed to enable further data exploration and interpretation of multiple levels of molecular data, providing visual context to the therapeutic potential of miRNAs.

2 Regulation of gene expression

'Every cell in an organism, with the exception of the sperm and egg cells, possesses the same set of genes. And yet a retina cell expresses genes to detect light and color, and a white blood cell expresses genes to fight infection. How can such different cells be created out of the same genetic blueprint?' - Siddhartha Mukherjee, *The Emperor of All Maladies* (2011)

Gene expression is defined as the use of a specific set of DNA (i.e. genes) to carry out cell maintenance and function [8]. Genes are transcribed into complementing messenger RNAs, which are then either translated into proteins or used to control gene expression (Figure 1). Thus, the set of genes *expressed* as RNA in a certain biological context reflects the current molecular state of cells and reveals potential pathological mechanisms underlying disease [6].

The world was baffled when, in 2001, the first publications of the human genome sequence [9, 10] revealed the number of protein-coding genes in humans was not much more than the number in worms (approximately 20,000). The expectation had initially been around 100,000 genes based on the approximate size of the genome divided by the average size of a protein-coding gene [9]. Surely the human was a functionally and anatomically more complex organism than the worm? The reasoning was then adjusted so that human complexity arose from 'doing more with less' — that is, the diversity seen in humans is the result of highly-intricate regulation of gene expression [5]. In line with this reasoning, the over 90% of the human genome designated *non-protein coding*, including functional elements regulating gene expression, is found to be biochemically active [7]. Continued improvements to genome sequence quality is seeing the number of protein-coding genes drop even further [11] and emphasizes the importance of expression regulation in higher organisms [12].

Dysregulated gene expression is associated to many complex diseases, from the large umbrella term of 'cancer' to a spectrum of neurological disorders. *Complex* refers to the multitude of integrated and interacting systems contributing to the disease [13]. Though a disease phenotype may be classified with a general level of accuracy, the diversity of the underlying gene expression may be vast. The varying degree of success in some cancer treatments, for example, is partly attributed to differences in how the expression of genes is regulated [14].

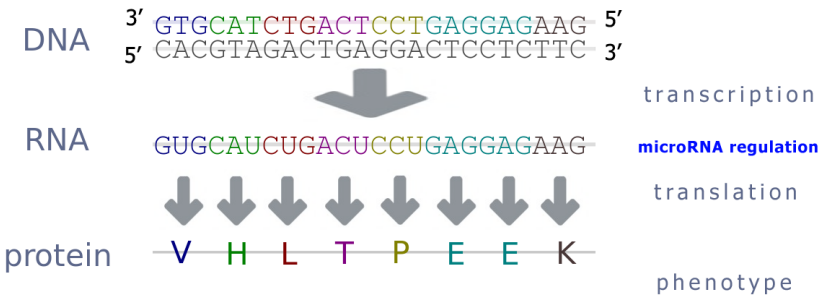

Figure 1: The flow of genetic information and the stage in which miRNA regulation occurs. Genes are expressed by first being transcribed into RNA and then translated into protein. MiRNAs generally inhibit gene expression at the post-transcriptional stage. Original image credit: Madeleine Price Ball, Creative Commons CC0 1.0 Universal Public Domain Dedication. *https://commons.wikimedia.org/wiki/File:Genetic_code.svg*

The Cancer Genome Atlas project [15] was launched in 2006 shortly after the Human Genome project completed (2003) with the goal of generating a comprehensive catalogue of the cancer genes and cancer-related mutations involved in tumourigenesis across a broad range of cancer types . The idea was that cancers, like other complex diseases, are so intimately associated with molecular changes that a comprehensive and systematic profiling of the molecular landscape across specific tumours would reveal important biological information with clinical implications. The primary medium for achieving its goals was next-generation sequencing technology (NGS), touted in its early years to be the key to discovering the cure for cancer. What it revealed, however, is that cancer is not a single type of disease but a whole class of diseases arising from different mechanisms in which uncontrolled cell growth and proliferation is achieved [16].

Shortly after the publication of the human genome sequence, it was proposed that RNAs with regulatory function form the primary control architecture enabling eukaryotic complexity and phenotypic diversity [5]. Non-coding RNAs represent a large portion of the genomic output and, unlike proteins, do scale-up in proportion to perceived eukaryotic complexity [2]. The discovery of sequence-specific gene silencing by RNA molecules, also known as the RNA interference pathway (RNAi), lead to the rise in prominence of miRNAs with the ability to regulate gene expression in an unanticipated range of biological processes [3].

3 MicroRNAs

3.1 Small molecules, big effect

MiRNAs are small, single-stranded RNA molecules approximately 22 nucleotides long with the ability to negatively regulate gene expression. On a post-transcriptional level, they are able to target specific gene transcripts through base-pairing to the mRNA 3' untranslated region (UTR) and either inhibit translation to proteins or activate mRNA degradation [17] (Figure 2). However, the initial discovery of miRNAs was not that of a class of functional RNAs but that of a single, regulatory RNA thought to be exclusive to nematodes.

Lin-4, the first known miRNA, was discovered in *Caenorhabditis elegans* in 1993 to have an important role in larval development [18]. The initial understanding of miRNA-mRNA target recognition came from the observation that *lin-4* could base-pair to multiple, conserved sites within the *lin-14* mRNA's 3' UTR [18, 19]. The second miRNA was *let-7*, discovered seven years later to also affect larval development in *Caenorhabditis elegans* [20]. Unlike *lin-4*, however, *let-7* was also

detected in several other species including humans [21]. This discovery marked the beginning of a field of study on a whole new class of negative, regulatory non-coding RNAs.

The known mechanisms by which miRNAs inhibit gene expression is similar in plants as in animals. In 2002, it was demonstrated that the 'seed' region of a miRNA sequence (that is, positions 2 to 7 or 8 of the 5' end) complemented short DNA response elements in *Drosophila* known to repress a host gene's expression post-transcription [22]. This suggested sequence complementarity was essential to post-transcriptional regulation by miRNAs. As more miRNAs were discovered, corresponding gene targets were generally recognized through base-pairing of the seed region of the mature miRNA sequence to such miRNA response elements (mREs) within the 3' UTR of mRNAs. How well a miRNA binds to a target gene's mREs influences whether the transcript is blocked from protein translation or degraded. Furthermore, the relative shortness of the required complementary sequence gives miRNAs the unique ability to simultaneously target and bind to multiple mRNAs, enabling an intricate level of gene expression regulation [23].

Given that miRNAs like *let-7* are highly conserved sequences found in numerous species, and that the first functional roles were essential to organism development with lethal effect if inhibited, it was no surprise when more miRNAs were eventually discovered with regulatory roles in other fundamental cellular processes such as apoptosis, differentiation, proliferation and metastasis [3]. Studies on RNAi, the naturally occurring gene silencing process by RNA molecules, further revealed miRNAs as an essential part of the functional unit which targets mRNAs [17].

In vitro overexpression of a single, unique type of miRNA has been shown to decrease levels of over a hundred mRNAs [12]. Similarly, deletion of a single, unique type of miRNA has been shown to result in a discernible phenotype change in both plants and animals [17]. However, low levels of expression change in target mRNA suggest a sophisticated role in coordinating and 'fine-tuning' expression: sharpening the borders of spatial or temporal gene expression domains, as in neural development, or to achieve target mRNA expression in an optimal range to ensure the silencing of unwanted signals [12].

Like gene expression, miRNA expression naturally differs according to a cell's developmental lineage and stage [24]. This attribute enabled profiling of particular pathological and physiological processes by miRNA expression, and qualifies them as candidate biomarkers [25]. Their ability to modulate gene expression further qualifies them as potential therapeutic targets in various disease treatments [26]. Elucidating the biological function of miRNAs with expression profiles corresponding to a particular condition is, therefore, a clinically relevant and active

area of miRNA research.

3.1.1 Genomic organization and biogenesis

MiRNAs are encoded throughout the genome as independent transcription units, as miRNA clusters, or as part of an intron of a host gene. MiRNAs are characterized into families of identical or closely related sequences, and as neighbors encoded in the same genomic cluster [27]. Approximately one-third of miRNAs are found in clusters with distances of less than 51 kilobases between them and are generally co-expressed [23].

Sequence families are defined by strong similarities in the mature miRNA sequence, specifically in the miRNA seed region (Figure 3a). These mature sequences can be found in multiple locations in the genome, are not necessarily clustered, and often differ from other sequence family members by one nucleotide. Consequently, functional analysis of individual members of a sequence family is challenging compared to single, unique miRNA sequences due to genetic redundancy [23]. For example, the miR-34 family is made up of six mature miR-34/449 miRNAs with copies in three genomic loci. The loss of miRNA expression from knocking out one functional locus can be compensated by the other loci [28]. These sequence features make miRNAs exceptionally robust to change but also challenging to study.

MiRNA biogenesis is a tightly-controlled cellular process summarized in Figure 2. Briefly, a miRNA gene is first transcribed by RNA polymerase II in the nucleus to produce a primary miRNA transcript folded into a hairpin-shape (Figure 3b). These miRNA hairpins are further processed by a pre-mRNA splicing complex (either the Drosha-complex or the Spliceosome) before export to the cytoplasm [27]. The Dicer complex cleaves the miRNA hairpin to release the mature miRNA sequence. The RNA-induced silencing complex (RISC) is activated when it is loaded with a mature miRNA sequence to identify target mRNA with [29].

3.1.2 Inhibition of gene expression

MiRNA-mediated gene silencing has the potential for broad impact on gene expression, with clinically relevant and practical application due to their role in the RNAi. MiRNA abundance in higher organisms likely confers an evolutionary advantage, such as introducing a robust layer of regulation to fundamental biological processes [12].

To effectively target and inhibit the translation of mRNA, mature miRNA sequences must first associate to an Argonaute protein in the RISC to activate the complex [26]. The miRNA then guides the complex to a specific mRNA target. Once a

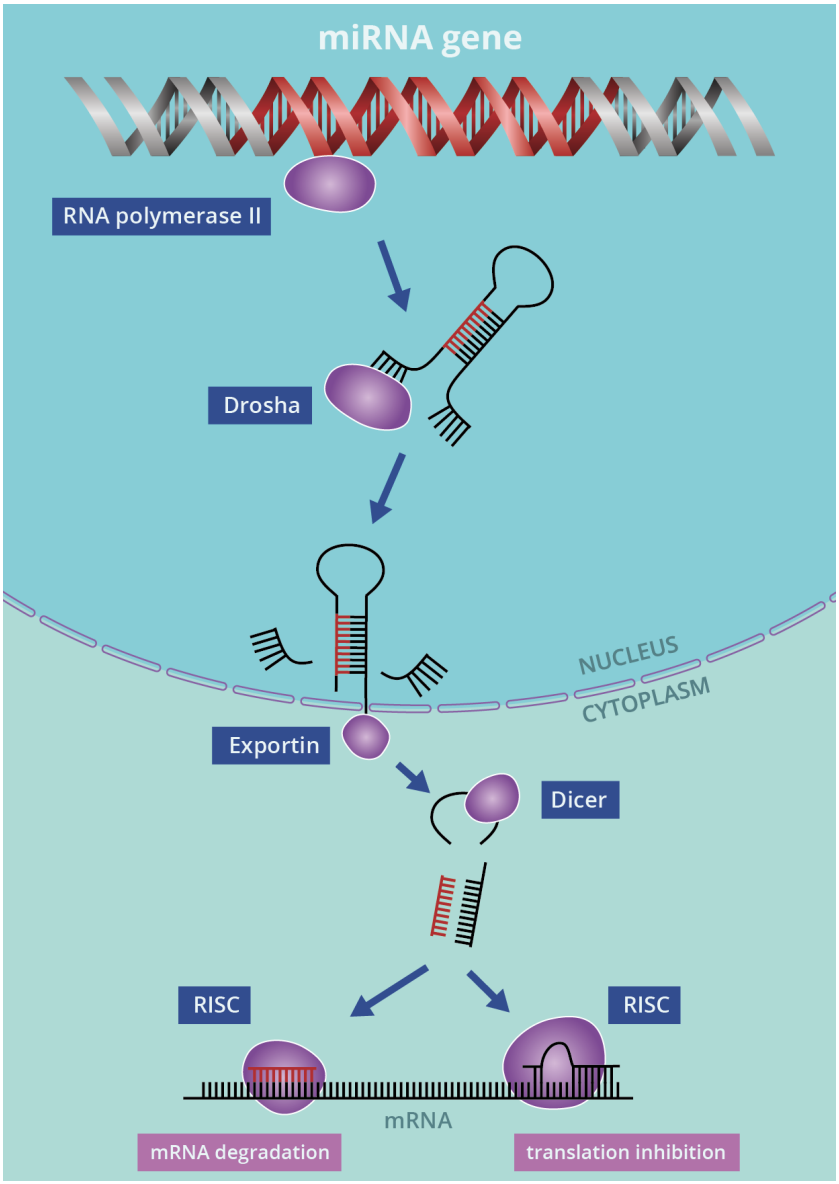

Figure 2: MiRNA biogenesis and function. MiRNA is initially transcribed in the nucleus. It forms a hairpin structure that is transported to the cytoplasm and further processed to produce a mature miRNA sequence. The sequence is loaded to an RNA-induced silencing complex and used to direct the complex to a target mRNA. The strength of the base-pairing between the miRNA and target mRNA determines whether the gene transcript is destroyed or inhibited from translation into a protein.

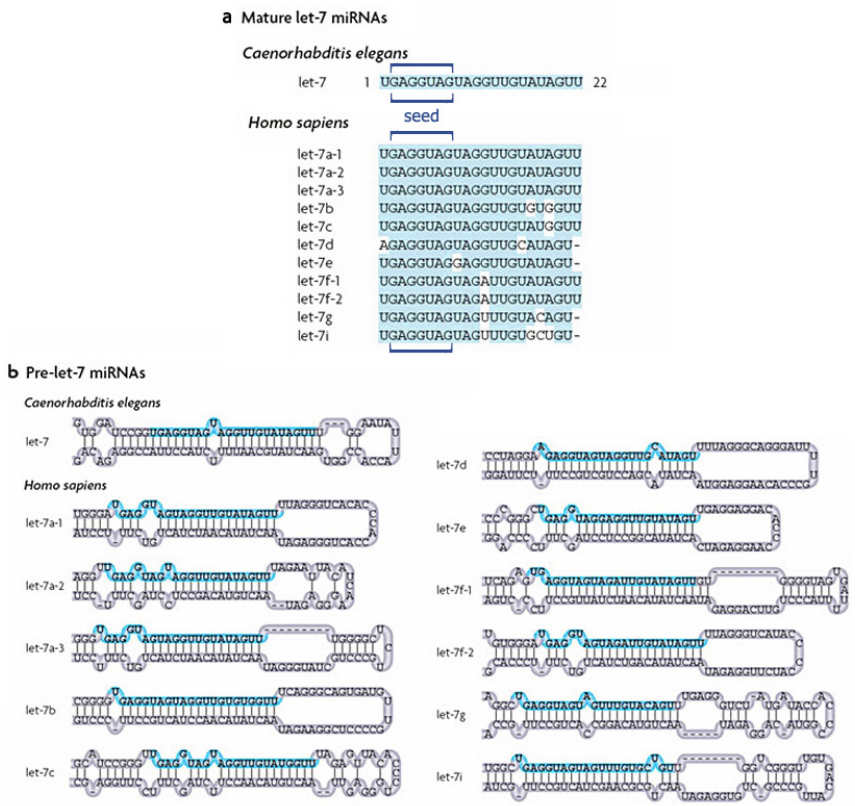

Figure 3: MiRNAs and miRNA genes. a) Mature sequences of the let-7 miRNA family members with seed region. b) Predicted hairpin structures containing the mature let-7 sequences (highlighted in blue). Reproduced with permission from Nature Publishing Group: *Micromanagement of the immune system by microRNAs* Lodish, H. et al. 2008 Nature Reviews Immunology, 8:120-130

target is found, the choice of translation inhibition or mRNA degradation is thought to be governed by the degree of sequence complementarity between the mature miRNA and the target mRNA [24].

The exact features and mechanisms of miRNA regulation on gene expression are still not sufficiently understood to accurately predict true molecular targeting and phenotypic effect [17]. What is generally known from extensive experimental validations is that degradation of a target mRNA occurs from the best (i.e. perfect or near-perfect) sequence complementarity, while weaker base-pairing results in translation inhibition [30]. Perfect sequence complementarity resulting in mRNA degradation is the most common case for miRNAs in plants. However, imperfect base-pairing is the most common case for miRNAs in animals. Still-

unconfirmed mechanisms are triggered such that both mRNA degradation and translation inhibition are possible [22]. It is thought that incorporation of sequence complementarity beyond the seed region to other features of the mRNA or RISC influences the method of suppression [23].

Though perfect and near-perfect sequence complementarity to the miRNA seed region has been and continues to be the most consistent feature used in the identification of target genes, growing evidence suggests it is not the only means miRNAs have to functionally target and regulate gene expression [31]. As a result, the search for true and actively targeted genes by miRNAs is not easy. Several sequence features are generally combined with seed sequence complementarity in computational algorithms to assist in the prediction of miRNA targets and are discussed in chapter 4.2.

3.1.3 MicroRNAs and complex disease

The ability of a single miRNA to control the expression of multiple genes means a single gene can also be regulated by multiple miRNAs. This makes it possible to coordinate a sophisticated level of regulatory control on disease pathways at multiple points [24]. In the decade following their initial discovery in humans, research has shown miRNAs with regulatory roles in numerous physiological, developmental and disease processes [3]. Over- or under-expression of certain miRNAs often correlate to a phenotype under investigation and a functional role is characterized by the pathways in which target genes are identified. In this way, miRNAs are nominated as candidate biomarkers and potential therapeutic targets in a number of complex diseases [26, 32].

In cancer, numerous studies have shown changes in miRNA expression correlate to tissue-specific and disease-specific conditions with direct effect in well-established tumour suppressor and oncogene pathways (e.g. TGFβ, TP53) [25]. Select miRNAs are highlighted in Table 1. Members of the miR-17-92 cluster have been identified in several cancers to target key genes in tumour suppressor pathways [14, 28]. Overexpression of the primary transcript encoding seven members of the cluster on chromosome 13 supports an oncogenic role for the cluster [33]. *TGFβ*-signaling is an important biological pathway for inhibiting cancer cell growth and down-regulation of a key pathway gene, *TGFBR2*, by members of the miRNA cluster has been shown to promote tumour cell growth [34]. Conversely, deletion of the primary transcript results in dramatically decreased levels of miR-17-92 and has been linked to developmental conditions such as microcephaly and digital defects [35].

Mir-663 targeting of *EEF1A2* has a tumour suppressive effect in pancreatic cancer

cells and has been shown to inhibit cell growth and invasion [36]. *EEF1A2* is normally not expressed or is in very small amounts in pancreatic tissue, but has been found at elevated levels in pancreatic cancer. Indeed, overexpression has been shown to promote cell growth and invasion in pancreatic cancer [37].

Cancer is often described as a complex disease because it cannot occur without the dysfunction of multiple, interacting biological systems. These systems consist of numerous molecular interactions where disruptions to the interactions with phenotypic effect are not always detectable on an individual molecular level [13]. Such interactions include the regulation of gene transcription and gene translation.

To better understand the combinatorial effect of dysregulated pathways, technology was developed to enable measurement of multiple levels of expression in a biological sample. This added depth improves identification of molecules with regulatory effect such as miRNAs [23].

Table 1: Select miRNAs with identified regulatory functions in associated disease.

miRNA	**Target gene**	**Associated disease**	**Reference**
miR-17-92 cluster	*TGFBR2, PTEN*	Lung cancer, breast cancer, autoimmunity	[33, 34, 35]
miR-663	*EEF1A2*	Pancreatic cancer	[36]
miR-10a/b	*HOXA3, HOXD10*	B-cell chronic lymphocytic leukemia, pancreatic cancer, breast cancer	[38]
miR-122a	*KLF6*	Liver tumour and disease	[39]
miR-let-7	*NRAS, KRAS*	Lung cancer	[40]

3.2 Technology to discover and measure microRNAs

Transcript-level analysis is an essential approach to characterizing and understanding the molecular functions underlying phenotype differences in biological samples. For decades, microarray technology was the most important and widely used approach for such studies [41]. It has only been since the publication of the human genome at the turn of the millennium that high-throughput sequencing of RNA (RNA-seq) emerged as a powerful and arguably better alternative [8].

Sequencing is the process of translating strands of DNA or RNA (i.e. reads) into their equivalent sequence of nucleotides. The technology was first developed by Frederick Sanger and colleagues in the 1970s and involved sequencing on a per-DNA strand basis. Though very accurate, the limited yield of a few thousand long

reads and the high costs (which, in the beginning, included manual base-calling) prevented most research projects from pursuing it as an alternative to microarray technology in facilitating transcript-level studies [42].

Microarray technology was developed in the 1990s as a hybridization-based approach to large-scale studies of gene expression that was both high-throughput and low-cost compared to traditional sequencing technology in terms of collecting sequence data [41]. Briefly, the array contains probes whose sequences represent particular regions of the genes to be detected. RNAs with complementary sequences hybridize to the probes and fluorescent is used to label the RNAs. The more RNA hybridizing to the probes, the stronger the signal intensity. Image acquisition of the array then enables measurement of the signal intensity or expression of the genes. [8]

There are a few limitations to microarray technology. Probes are designed to represent a given set of known genes under investigation, but microarrays may be measuring only the portion of a known gene corresponding to the probes and not the actual sequence of all transcribed RNAs [8]. The range of expression measured is also limited due to the presence of background noise and signal saturation. Background noise is introduced by imperfect probe hybridization to semi-complementary RNA sequences. In contrast, signal saturation is introduced when the number of complementary RNA sequences outnumber available probes.

The push to sequence the human genome brought with it the advent of the first high-throughput or 'next generation' sequencing (NGS) platform. The first decade of the twenty-first century saw rapid development of sequencing technology result in a huge drop in cost and a substantial increase in sequence yield [43]. Massively parallel sequencing — the ability to simultaneously sequence several hundred DNA sequence fragments — enabled the substantial leap in data volume as well as data generation speed, while protocol innovations enabled sequence production on a genomic, transcriptomic, and proteomic level [42].

NGS as a medium for expression analysis can be roughly summarized as follows: RNAs are initially fragmented into short strings and reverse-transcribed into complementary DNA (cDNA). Adaptors are added to the cDNAs to enable library amplification and sequencing [8]. Amplification achieves high signal intensity that enables detection of otherwise low-expressed molecules. Rather than hybridize to an array, the sequenced reads can then be mapped to a reference genome. The number of reads aligned to a gene are counted to give a digital measure of the expression levels in the sample under investigation.

Several comparisons of microarray and NGS technology have been published [8] and include Pub IV. While comparisons have highlighted strong concordance

between microarrays and sequencing in measures of both absolute and differential expression, the general consensus is that NGS:

- can reconstruct transcripts at a single-base level. This enables expression quantification and analysis of both known and novel RNA sequences.
- enables *de novo* genome assembly of non-model organisms.
- has the advantages of massively parallel sequencing which include low background noise and high technical reproducibility. [8]
- is preferable to microarray for differential expression analysis because of its high coverage of the genome and detection of weakly expressed genes [44].
- is not limited by signal saturation. Counting sequenced reads achieves a broader dynamic range and a more sensitive measurement of transcript abundance than signal intensity.
- is not as straightforward to process and analyze as microarrays due to the variety of protocol differences that need to be accounted for in data preprocessing and normalization.

For these reasons, and the continued reduction in cost for large-scale studies, NGS has since replaced traditional sequencing and microarray as the preferred technology for expression profiling and transcriptome-level studies. This work focuses on analyzing miRNA expression profiles using NGS technology.

The volume of biological data now being produced with NGS technology has brought new opportunities to study biological complexity but has also raised challenges in both analysis and data management. Before any computational analysis on the large volume of data can be implemented to identify interesting molecular features, computational resources for large-scale data processing and storage must first be obtained.

While standard processes and tools have been established for microarrays, they are not directly applicable to NGS data because of the fundamental differences in data collection and measurement. Microarray signal intensities, for example, are continuous measures that follow a log-normal distribution and so are not directly comparable with NGS read counts, which are non-negative and discrete measures [45]. As such, different normalization approaches must be applied.

Computational tools have been developed and/or updated to address the new technical and biological challenges of NGS technology to perform analysis comparable to established microarray standards. Tools have also been developed and updated to take advantage of the opportunities in NGS technology, such as quantifying gene isoform expression [42, 46]. In the case of miRNAs, sequenced reads facilitate the prediction of likely target genes through the discovery and identification of complementary sequenced regions.

Discoveries made with NGS technology have lead to the further development of RNA-binding technology such as cross-linking ligation and sequencing of hybrids (CLASH) [47], which makes it possible to identify direct RNA-RNA hybridization. However, this work is focused on the use of NGS technology to facilitate expression-based analysis, so more specialized technologies are beyond its scope.

4 Processing and analyzing sequenced microRNA data

Studies involving RNA-seq experiments are generally guided with specific questions and research goals. The diversity of sequencing protocols and experimental designs possible with RNA-seq facilitates a range of sequence-based studies, but prevents development and implementation of a universal RNA-seq processing and analysis protocol [42]. However, most RNA-seq studies do generally include the following computational steps: sequence preprocessing, read mapping to a reference genome or assembly, expression quantification, normalization and analysis.

Preprocessing largely consists of NGS adaptor removal from raw sequenced reads and the trimming of low-quality portions of the sequence. To measure transcript abundance, processed reads are then mapped to a desired reference genome based on high sequence complementarity. Expression quantification is counting the number of reads mapping to genomic regions of interest (annotated to genes, for example) as either read counts or reads per kilobase per million mapped reads (RPKM). The latter normalizes counts for total read length and the number of reads, but is not sufficient normalization alone to put sequence expression across samples on a comparable scale for analysis [44]. Furthermore, the choice of normalization depends on the analysis in question, and for differential expression the normalization has been incorporated into the analysis process [8, 48]. These are discussed further in Methods.

4.1 Computational workflow

A workflow is an orchestrated sequence of events enabling systematic organization of tools and resources to produce a desired outcome. Computational workflows can be further automated and organized such that each step is a self-contained module made of interchangeable and reusable component parts [49]. This allows for robust customization of a workflow to better fit the characteristics and expectations of a research project.

With the rapid evolution of RNA-seq technology and consequent absence of an established standard for processing and analyzing RNA-seq data, modularity is an

essential feature in an RNA-seq workflow. Modularity allows for straightforward incorporation of tools and methods for each step in the RNA-seq analysis process. Having multiple options is necessary because the tools and methods developed at each step of the process have been shown to be good for some types of RNA-seq data and experimental design but not all [45, 50]. The diversity in computational tools reflects the diversity in RNA-seq technologies, which themselves are constantly developing and improving. Ideal workflows must thus be able to provide a standard software framework, but have the robustness to accommodate tool upgrades and innovations in RNA-seq technology processing and analysis [51].

The use of multiple levels of biological data has the advantage of addressing research questions about biological complexity, but the disadvantage of requiring additional management of resources to ensure efficient processing. Computational workflows that scale-up well (i.e. utilize parallelization and batch processing) are thus essential for miRNA studies, which are rarely performed without complementing mRNA and/or gene data to facilitate functional analysis.

Anduril [52] is an open-source workflow engine that provides the necessary infrastructure and features (outlined in Pub I) for reliable and scalable data analysis. Briefly, a workflow in Anduril automates the process and analysis of RNA-seq data as a defined sequence of components. A single component can perform simple, reusable tasks such as table filtering or more complex and specific tasks such as differential expression analysis with several R packages. Tools utilized within components are usually written in a command-line executable language (e.g. Bash, Java, R, Python, Perl and MATLAB) and unified under a simple, Java-based language (AndurilScript) to enable component-based workflow bulding. Features of Anduril further enable users to define parameters to better fit an experimental design, optimal use of computational resources (including parallelization of tasks), and fast re-execution.

It is difficult to compare the performance of tools across the range of possible RNA-seq applications largely due to differences in performance optimization and development for specific research questions [53, 54]. Comparison of tools is also beyond the scope of this work. General features of the NGS data, however, ensure a level of similarity exists between approaches as to how the data is used. Two such approaches that use sequenced data for miRNA studies are novel miRNA discovery and miRNA target prediction, which are discussed below.

4.2 Novel miRNA discovery

NGS technology not only facilitates improved measurement of transcript abundance but also detection of transcript variety. While some miRNAs are conserved across

multiple species such as *let-7*, some are species-specific such as *lin-4*. The discovery of a potentially novel, species-specific miRNA is done through analysis of the possible secondary structure of an expressed sequence that maps to the genome but not to any known miRNA sequence. The ability for the genomic area around the sequenced transcript to have a secondary structure that forms a hairpin-like shape is a key feature of miRNAs [19].

The discovery of potentially novel miRNAs is not a standardized process but methods do tend to incorporate the presence of a biochemically stable hairpin secondary structure in the vicinity of the proposed mature miRNA [55]. The hairpin shape seen in Figure 3 is only possible with the presence of a complementary sequence to the proposed mature miRNA in the mapped area of the genome. The ViennaRNA package [56] is a RNA secondary structure tool that predicts likely paired sequences by creating base-pairing probability matrices for a particular base pair. While the presence of a putative mature miRNA sequence implies the existence of a hairpin, the reverse is not always true [29].

Cross-species conservation is another determining factor for considering a putative novel miRNA sequence. However, it risks missing non-conserved miRNAs. The more distant two species are in phylogeny, the less likely conservation-based definitions are to facilitate the identification of new miRNAs. It is estimated that 7% of human miRNAs are species-specific [23].

Machine learning techniques have been applied to the discovery and analysis of putative miRNA sequences based on features inferred from existing miRNAs [57]. These approaches assume that currently known miRNA sequences are representative of all existing and yet-to-be discovered miRNAs and thus contain all known and unknown features to identify a miRNA.

Novel miRNA discovery is generally performed to identify previously unknown miRNAs in a tissue- or species-specific context. This is particularly useful in the study of disease, where mutations in key regulatory sequences can not only affect normal function but also cell phenotype [23]. Methods to identify putative novel miRNAs are generally based on sequence feature identification and analysis, with features defined by already known miRNA sequences. However, due to the yet-unknown factors determining their tissue-specific expression and inherent genetic redundancy, computational methods alone are insufficient to claim discovery of a novel miRNA. Further experimental analysis is needed to identify active function and provide evidence of a relevant biological effect.

4.3 Prediction of miRNA targets

To characterize the biological function of a miRNA, one must discover its target genes. The primary approach is guilt-by-association: coding transcript expression that rises or falls inversely to a specific miRNA's expression change are identified as potential targets [31]. This provides basic evidence that the miRNA likely inhibits the gene's expression and has a role in regulating the gene's annotated pathway. Such a functional hypothesis can then be experimentally tested.

While technology exists that enables the discovery of miRNA targets through the direct binding of a mature miRNA to a mRNA transcript [29, 47], it is neither resource-efficient nor financially feasible to experimentally test all possible miRNA-target gene combinations for true expression inhibition. Therefore, computational tools are often used to identify the *most likely* and context-specific target genes for experimental validation. In other words, prioritizing likely target genes by the presence of miRNA regulatory features and by annotated functions important to the disease and/or biological context under investigation.

Available programs developed for target prediction tend to require some degree of sequence complementarity to the miRNA seed region and favorable free-energy in the miRNA-target duplex. Free-energy estimations are measurements of the accessibility of a nucleotide binding to a complementing nucleotide according to the RNA secondary structure. As it involves the analysis of RNA secondary structure, the measurement is a defining criteria for both novel miRNA discovery and miRNA target prediction [58].

Commonly used criteria for predicting and filtering miRNA targets are shown in Figure 4 and include:

- MiRNA seed pairing to a complementary mRE in the 3' UTR of a putative target mRNA, with identification allowing for up to two mismatched bases and additional complementarity outside of the seed region to compensate miRNA-target gene binding stability [31].
- The identification of multiple, close-proximity mREs for the same miRNA in the 3' UTR of a putative target mRNA. The presence of multiple potential regulatory elements increases the likelihood of a miRNA-mRNA interaction occurring [30].
- Conservation of a target gene's mREs in related species. Conservation suggests an evolutionarily conserved interaction. When conservation is not possible, programs like *Targetscan* [59] account for species-specific miRNAs by putting more emphasis on sequence context (i.e. base complementarity outside of the seed region).

- Availability of the mRE and the estimated thermodynamic stability of the proposed RNA interaction. Minimum free energy (MFE) is a measure of how stable the predicted miRNA-mRNA duplex is [31]. *MiRanda* was the first freely available target prediction algorithm and it used MFE to improve prediction accuracy [60]. *PITA* also calculated the stability of the predicted miRNA-mRNA duplex but compared it to the stability of the local 3' UTR region of the mRE [61]. The idea was that if the local region of the mRE was more stable, then it would prevent the miRNA-mRNA duplex from forming.
- Filtering for inversely correlated miRNA and target mRNA expression data. Filtering has often and successfully been applied to increase target prediction accuracy because it is independent of sequence analysis [31]. However, it cannot distinguish between a direct and indirect target like sequence analysis can. Also, because the presence of a miRNA does not always result in observable changes to mRNA expression, filtering reduces the set of predicted targets to those notably affected [28].
- Filtering for enrichment of biological pathways by predicted target genes. Filtering predicted targets to those in common pathways improves context specificity of the experiment [31]. MiRNA regulatory function is thus defined by the similarity of functions and pathways annotated to the set of predicted target genes.

In general, a combination of target prediction algorithms has shown to produce better results than target prediction algorithms used separately. The combination expands the hypothesis space to incorporate the many possible features of miRNA regulation [31]. The underlying idea is that there is more than one way for a miRNA to target genes and silence their expression. The availability of multiple target-binding features is how a miRNA is able to regulate the expression of hundreds of genes simultaneously. Each target prediction algorithm has been optimized to capture different combinations of miRNA target-binding features and thus achieves different levels of sensitivity and specificity[1] [31]. Target genes predicted by more than one algorithm, therefore, gain confidence as likely miRNA targets.

Predicted targets are only relevant to a specific phenotype if they are expressed in the tissue under investigation. This is an issue not addressed by most prediction algorithms and is why filtering of results is necessary. A current and actively-maintained list of miRNA-target interactions with experimental support is a valuable resource for identifying the most likely miRNA target genes in a biological context. The miRTarBase database [62] is one such resource accumulating miRNA-target interactions by manually curating literature and using text mining to filter research

[1] Specificity here is the ability to discriminate between an intended target and non-targets. That is, the miRNA will bind to the former but not the latter.

articles related to functional studies of miRNAs.

To summarize, the prediction of a miRNA's targets is the guilt-by-association approach to characterizing its biological function in a tissue-specific context. Because the exact mechanisms in which miRNAs identify their target genes and inhibit their translation are still unknown, computational methods are needed to predict their most likely target genes. Existing tools and algorithms focus on sequence analysis to model different combinations of RNA-binding features. Additional filtering for tissue-specific expression profiles lends evidence of notable regulatory effect, while pathway annotation identifies gene subsets with interesting roles associated to the tissue under investigation. Thus, miRNA target prediction is a computational process that identifies a priority set of molecules for further investigation. Like novel miRNA discovery, biological expertise and experimental analyses are still needed to ascertain molecular function.

Even with biological expertise, it can be a challenge to translate statistically significant results into meaningful and actionable clinical insights [63]. MiRNA function is primarily defined by its interaction with target mRNA and effect on associated biological pathways. Data visualization offers a straightforward and intuitive approach to interpreting such interactions that does not require additional understanding of the algorithms underlying the data [64]. It offers a convenient and effective means to assess the potential relevance of a priority set of molecules through their expression data [65]. Modern visualization technology enables data exploration, automatic generation of, and efficient manipulation of diagrams from quantitative values [66]. Furthermore, data exploration incorporates a human computing approach to an analysis workflow, harnessing our capacity to visually detect patterns to improve hypothesis generation [64].

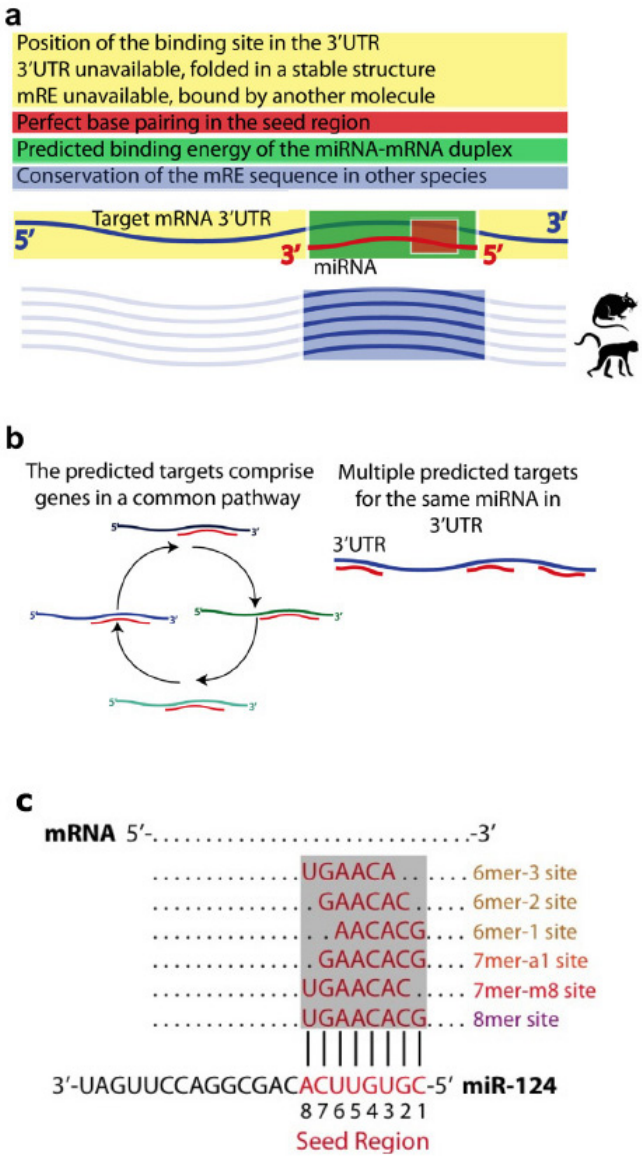

Figure 4: Commonly used features in miRNA target prediction. a) Regions of the target mRNA 3' UTR and mature miRNA containing miRNA-target mRNA binding features. Conservation of the mRE in other species is used in target prediction, while conservation of the mature miRNA sequence in other species is used in novel miRNA discovery. b) The likelihood of a target increases if genes with similar function or within the same biological pathway are also targeted, and if multiple mREs are found in the target's 3' UTR. c) Base-pairing of the seed-region is the primary feature of target prediction algorithms but is applied in various ways to allow for up to two mismatched bases. Reproduced with permission from Elsevier: *Refining microRNA target predictions: sorting the wheat from the chaff* Ritchie, W. and Rasko, J.E. 2014 Biochem. Biophys. Res. Commun. 445, 780-784.

5 Aims of the study

The aim of this work was to develop and use computational tools to better facilitate miRNA research in biomedicine. Emphasis was placed on the end goal of identifying miRNAs with potential as biomarkers and therapeutic targets.

The specific research objectives covered in this study were:

1. Development of a workflow to accurately process and analyse miRNAs and complementing mRNA data derived from RNA-sequencing technologies. The workflow provides a convenient implementation of a suite of computational tools to produce standardized and reproducible results. (Pub I)
2. Application of the workflow to miRNA and complementing mRNA in biological studies, with functional analysis identifying candidate miRNA biomarkers and therapeutic targets. (Pub III and a portion of Pub IV)
3. Adaptation of a visualization method to meaningfully depict interactions between multiple levels of molecular data, such as miRNA co-regulation of genes in a biological pathway, and to facilitate interpretation of findings to clinically relevant implications. (Pub II)

6 Materials and methods

Biological material, third-party target prediction databases, and computational methods used are summarized in this section with further detail found in the publications.

Table 2: Biological data by publication. Asterisk (*) denotes publications using original data.

Publication	Sample	Data type & platform
Pub I	Breast tissue. Case 1: mRNA from 17 patient tumour and 3 normal breast organoid samples. Case 2: mRNA from 129 tumour and 15 germline samples, miRNA from 133 tumour and 16 germline samples.	Case 1: Open total-RNA sequence data, Illumina HiSeq 2000 (GSE52194). Case 2: TCGA generated Level I poly(A)-extracted mRNA and small RNA sequence data, Illumina Genome Analyzer II.
Pub II	OVCA tissue. miRNA and mRNA from 324 patient primary tumour samples, subset of 32 good prognosis and 32 poor prognosis.	Open miRNA and mRNA expression data generated by TCGA. mRNA microarray, Affymetrix U133A. miRNA microarray, Agilent. [67]
Pub III*	DLBCL tissue. miRNA and mRNA from 7 patient fresh frozen samples, matched primary and relapsed tumours.	small RNA and total RNA sequence data, Illumina HiSeq 2000 (GSE69810)
Pub IV*	C57BL/6J mouse brain tissue. miRNA from 6 snap frozen samples, matched frontal cortex and hippocampus.	miRNA expression, Illumina Cluster Station and Genome Analyzer II (GSE27979). miRNA expression, Affymetrix GeneChip® miRNA array (GSE27891).

6.1 Biological data

Table 2 briefly summarizes the sample data analyzed in each publication. For mouse samples and high-grade serous ovarian cancer (OVCA) tumour samples, analysis began with fully processed and quantified expression data. For diffuse large B-cell lymphoma (DLBCL) and breast cancer samples, sequence data was obtained for us to process ourselves.

Affymetrix GeneChips® miRNA microarray was used in Pub IV and is based

on 609 known mouse miRNAs in version 11 of the miRNA database, miRBase [68]. Sequenced miRNA and mRNA transcripts were aligned to the human genome (NCBI38v76) in Pub I and in Pub III. MiRNA sequences were then annotated to known human miRNAs in version 21 of miRBase, which corresponds to the same genome reference. TCGA-generated breast cancer and OVCA data used in Pub I and II were obtained from the data portal[2]. Further sample descriptions are found in the publications.

In Pub II, prior work by Yang *et al.* [69] identified 219 genes and 19 targeting miRNAs as a master miRNA-gene regulatory network associated to poor prognosis in OVCA mesenchymal subtype. Microarray expression for these identified miRNA-gene pairs were obtained for further expression correlation and pathway impact analysis. From the original TCGA project of 324 patient samples with clinical information, we also identified the top 10% of patients with the most days to tumour progression recurrence as having *good prognosis* and the lowest 10% of patients with the least days to tumour progression recurrence as having *poor prognosis* for a total of 64 samples. Differential analysis was performed on microarray gene expression [70] and was followed by pathway impact analysis. Predicted targeting miRNAs were then identified for the set of DE genes enriched in significant pathways (with false discovery rate $\leq 10\%$) and expression correlated across the 64 samples.

6.2 Predicted miRNA targets database

Putative miRNA-mRNA pairs were downloaded from databases listed in Table 3 and annotated to miRNAs and mRNAs with significant, inversely related expression profiles. This filtered and prioritized a subset of potentially interesting miRNAs and their putative target genes in a particular experiment [71, 72, 73]. Each algorithm represents the implementation of similar target-binding criteria with varying emphasis on each criterion. A combination of algorithms thus captures the general principles of miRNA-target gene interactions [31].

In Table 3, *seed pairing* includes imperfect sequence matches of the seed region, perfect 8 nucleotide seed match, allowance of guanine base-pairing with a uracil instead of a cytosine (G:U wobble), and in the case of *RNA22*, motifs from known miRNA sequences and the complements used to identify mREs. *Thermodynamics* includes MFE of the seed pair and mRE site accessibility to targeting. *mRE region* covers all detected sequence features in the environment of the mRE, including

[2]http://cancergenome.nih.gov

Table 3: MiRNA-target mRNA prediction databases.

Database	Publication	Algorithm approach
TargetScan	I, III, IV	Seed pairing, conservation, mRE region [59].
Microcosm (MiRanda)	I, III, IV	Seed pairing, conservation, thermodynamics [60].
RNA22	IV	Seed pairing & thermodynamics [74].
RNAhybrid	IV	Seed pairing, thermodynamics, mRE region [75].
mirDB	IV	Machine-learning, seed pairing, conservation, thermodynamics [76] .
PITA	I, III	Seed pairing, conservation, thermodynamics, mRE region [61].
DIANA-microT	I, II, III	Seed pairing, conservation, thermodynamics, mRE region [77].
mirTarBase	I, III	Manually curated, experimentally validated targets [62].

additional sequence complementarity, thermodynamic stability, and presence of multiple mREs [78].

6.3 RNA-sequence processing and analysis

All biological data used in the publications of this work went through similar processing and analysis steps, either performed by us or a third-party. These steps are generalized as: preprocessing, read mapping to a reference genome or assembly, expression quantification, normalization, analysis, and miRNA-mRNA integration (Figure 5).

To preprocess raw sequenced reads in Pub I and III, quality checks were first performed to determine required trimming parameters. Read statistics, adaptor trimming, and post-trimming quality control then followed. Samples with poor quality scores or with insufficient number of reads surviving adaptor and quality trimming were excluded from further processing.

For read mapping, the choice of sequence aligner depended on the type of sequence and optional feature discovery. The STAR aligner [79] was used in Pub I to map RNA to the human transcriptome and identify potentially novel transcripts. Bowtie [80] was used in Pub I and III to map small RNA to known miRNA transcripts. It was also configured to identify reads that mapped to the human genome but not to any known miRNA transcripts for the purpose of novel miRNA and other small RNA discovery.

Small RNA annotations were extracted from Ensembl general transfer format (.gtf) files and referenced for read quantification. MiRNA annotation files in general feature format (.gff3) were also downloaded from miRBase [68]. Though they share a lot of the same genomic information, differences in miRBase's annotation format compared to Ensembl makes it incompatible for use with the expression quantification tool HTSeq [81]. SePIA includes a transitional step to reformat the .gff3 file to resemble that of an Ensembl .gtf file. This enabled expression quantification on a mature miRNA level not possible with Ensembl transcript annotation.

The estimated expression of a gene, gene transcript, mature miRNA, or miRNA transcript is quantified from mapped reads. However, count-based expression generally require further normalization to be informative. The three methods implemented in SePIA are counts-per-million with edgeR [82], library-size factor scaling with DESeq [83], and upper-quartile normalization [84].

Analysis of processed reads, mapped reads, and/or expression data differs depending on research goals. Differential expression analysis is the most common analysis and was performed for this work in Pub I, III and IV to identify candidate miRNA biomarkers. Novel miRNA discovery and miRNA-mRNA integration in Pub I demonstrated SePIA workflow capabilities. Novel miRNA discovery and miRNA-mRNA integration in Pub III identified miRNAs and their putative targets with statistically significant potential in determining post-treatment prognosis. Expression analysis and miRNA-mRNA integration was also performed in Pub II and IV to identify candidate miRNAs for further functional analysis.

The predicted targets of differentially expressed (DE) miRNAs were queried from the target prediction databases described in Table 3. Expression profiles were then extracted for each of these putative targets and correlated with the corresponding miRNA expression. Target genes were then filtered for moderate to high expression anti-correlated to a targeting miRNA. This produced a list of miRNA-target gene pairs whose expression profiles showed a significant, inverse relationship supported by at least one target prediction database. Pathway impact analysis identified the biological pathways enriched by the putative miRNA targets, inferring likely regulatory roles for miRNAs in distinguishing tissue phenotype. MiRNA-mRNA integration results were visualized with Director in Pub I, II, and III to aid hypothesis generation and guide further functional analysis.

To ensure reproducibility of results, each publication work is implemented as a custom workflow. Figure 5 outlines the general steps used in each workflow and Table 4 lists the software. SePIA is the workflow template that contained all the essential steps for the processing, individual analysis, and integrated analysis of

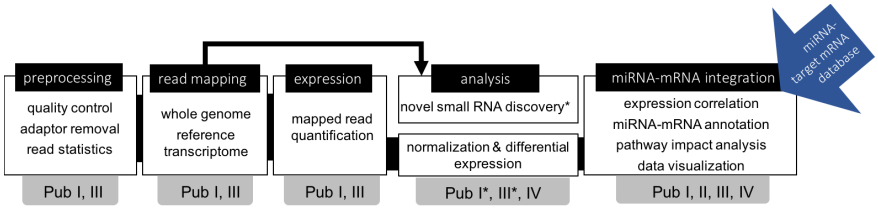

Figure 5: Overview of the computational workflow implemented for the studies in this work. Steps are labeled with the publications they were implemented in. An asterisk (*) denotes publications where novel small RNA discovery was performed. Target prediction databases are described in Table 3. Software tools utilized at each step are described in Table 4.

RNA and small RNA data. The workflow was used in Pub I and III for both RNA-seq and small RNA-seq data, but also in yet unpublished work for just RNA-seq or just small RNA-seq data. It is the successor of a rudimentary workflow used in Pub IV, which consisted of expression analysis and miRNA-mRNA integration. Pub II was limited to miRNA-mRNA integration with emphasis on data visualization.

6.3.1 Differential expression analysis

In Pub I, III and IV, DE miRNAs were identified for further analysis. At its simplest, detection of differential expression is the application of a test statistic to identify molecules with expression differences between experimental groups exceeding the predicted range of variability [8].

Microarray technology has a well-established method for normalizing and identifying DE molecules that was implemented in Pub IV. The R package affy [85] took the raw, probe level data from the Affymetrix microarray and quantified expression between samples as log-intensities using the robust multi-array average (RMA) measure. The package limma [86] then performed differential expression analysis using a t-test to compare the log2-fold change of a miRNA's expression to that of all other miRNAs.

Differential expression analysis of RNA-seq data generally consists of two steps: estimation of model parameters from data and tests for differential expression [45]. Parameters take into account that transcriptome abundance is affected by both biological and technical variation across samples (e.g. read failure in low-count samples) [48]. The Negative Binomial distribution is a discrete probability distribution suitable for count-based expression data [8]. The two parameters of the distribution encode mean and dispersion of counts. Dispersion describes how much the variance deviates from the mean count of a gene across samples. Thus, overdispersion is when the variance of gene expression across multiple biological

replicates is larger than its mean expression values [87]. The relationship between the variance υ and mean μ is defined as,

$$\upsilon = \mu + \alpha\mu^2$$

where α is the dispersion factor.

Pub I and III utilized two of the most established methods, the R packages DESeq [83] and edgeR [82] for identifying DE miRNAs. Pub IV utilized edgeR for sequence-derived miRNA count expression. The two methods use the Negative Binomial model and implement similar strategies to reduce the inherent bias of transcriptome abundance [88, 89]. As a result, they have been shown to produce similar sets of DE genes from large enough datasets [45, 48, 50].

Table 4: Software implemented in the workflows. Software are marked with the publications they were used in. *Italicized software* are not part of the SePIA workflow and were run separately.

Module	Component	Software	Reference
Preprocessing	Adaptor and quality trimming	FastX-Toolkit^{PubI,III}	[90]
		Trimmomatic^{PubI,III}	[91]
	Quality statistics	FastQC^{PubI,III}	[92]
Read mapping	Align sequences to a reference	Tophat^{PubIII}	[93]
		Bowtie^{PubI,III}	[80]
		STAR^{PubI}	[79]
	Alignment sorting and conversion	SAMtools^{PubI,III}	[94]
		Picard tools^{PubI,III}	[95]
	Alignment statistics	RNA-SeQC^{PubI,III}	[96]
		RSeQC^{PubI,III}	[97]
Expression	Mapped reads quantification	HTSeq^{PubI,III}	[81]
		Cufflinks^{PubI,III}	[98]
Normalization and Analysis	Variant calling	Bambino^{PubI}	[99]
	Transcript differential expression	Cuffdiff^{PubI}	[100]
	Normalization and differential expression	*affy*^{PubIV}	[85]
		limma^{PubIV}	[86]
		SAM^{PubII}	[70]
		DESeq^{PubI,III}	[83]
		DESeq2^{PubI}	[88]
		EdgeR^{PubI,III}	[82]
	Novel miRNA discovery	miRanalyzer^{PubI,III}	[57]
MiRNA-mRNA integration	MiRNA predicted target database query	sqldf^{PubI,III}	[101]
	Pathway impact analysis	SPIA^{PubI,II,III}	[102]
		Ingenuity® Pathway Analysis^{PubIV}	[103]
	Data visualization	*Director*^{PubI,II,III}	[Pub II]

7 Results

This dissertation presents the following main results: a workflow to process and analyze small RNA- and RNA-seq data (Pub I), application of the workflow – in a partial, rudimentary form (Pub IV) followed by a comprehensive form (Pub III) – to characterize the role of miRNAs in different tissues, and a visualization approach to further link and interpret results from multidimensional data (Pub II). Contributions from each publication are further summarized in Table 5.

Table 5: List of contributions in the dissertation.

Publication	Type	Summary
Pub I	Method	Workflow design and templates to process and analyze RNA and small RNA-seq data.
Pub II	Method	R package for visualizing multidimensional data.
Pub III	Biomedical	Profiled miRNA expression in matched, primary and relapsed DLBCL and functionally analyzed select miRNAs with complementary mRNA data.
Pub IV	Biomedical	Analyzed and compared microarray to small RNA-seq expression data for miRNAs. Identified DE miRNAs in both platforms and their putative target genes. Performed pathway analysis on predicted gene targets of select miRNAs to identify co-expression relationships and support hypotheses of miRNA function in brain tissue differentiation.

7.1 Workflow for small RNA- and RNA-seq

Pub IV showed that RNA-seq had the benefit of better detection overall of miRNAs than microarray, but also revealed novel computational challenges: the lack of gold-standard alignment, processing, normalization, and analysis methods; resource

availability; frequent software updates; diversity in experimental designs and methodology. A computational workflow was thus needed to address these challenges and to facilitate the individual and joint analysis of multiple forms of RNA-seq data. In this work, we focused primarily on small RNA (< 200nt), total RNA ($\geq$ 200nt), and poly(A) derived RNA-seq data.

SePIA was developed to incorporate a collection of tools representative of widely used and established methods for RNA-seq. It succeeds the workflow of Pub IV, where expression analysis was performed on already processed data. The quality of analysis results, however, depends on knowledge of how the data was processed and if the process can be replicated. So where it was possible to start the workflow from unprocessed sequenced reads (e.g. Pub I and III), parameters were better optimized to produce thorough and robust analysis results.

SePIA contains the following features of an ideal RNA-seq workflow: the organization and convenient implementation of several tools in sequence, documentation and reporting at each step of the workflow to ensure reproducibility, and the ability to specify computational resources and component parameters to fit an experimental design.

One of the challenges of high-throughput workflows is the amount of time, resources, and computational proficiency required to simply implement all the necessary tools [51]. SePIA has a convenient solution that uses Docker [104] to provide a portable and easy to load container from which workflows can be run with all dependencies pre-installed. The container ensures all essential software perform safely and reliably within a workflow, while keeping the host system clean of excess libraries. Users of SePIA can thus skip the daunting task of software installation and jump right into configuring the template pipelines for their data. HTML reports for each module and documentation in the form of log files enable reproducibility of results.

SePIA and its pipelines have been used in the cited publications and several ongoing biomedical studies involving mice, rat, and human samples.

7.2 Expression analysis identifies miRNAs with putative roles in tissue differentiation

MiRNA expression profiles in each publication were used to answer standard questions such as ‘what miRNAs are expressed?’ and ‘are there miRNA profiles characterizing clinical differences in samples?’ Differential expression analysis provides an answer to the second question by identifying a subset of molecules with potentially interesting roles in tissue-specific biological pathways associated

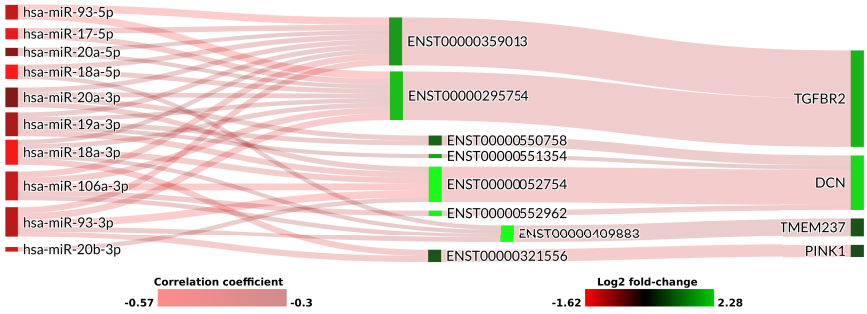

Figure 6: Target genes of the mir-17/92 cluster and paralog clusters in the TGFβ signaling pathway of Pub I. Target transcripts were selected based on a minimum log2-fold change of 0.5 between tumour and normal breast tissue. Node colors represent expression fold change between normal and tumour breast tissue samples, with higher expression in tumour tissues highlighted in red and in normal tissues highlighted in green. Relationships between connected miRNA-target transcript pairs are shaded based on correlation coefficient values. Connections between transcript and gene represent average correlation values of contributing transcripts and their regulating miRNAs.

to complex disease. In this work, miRNAs were identified that had significantly altered expression in healthy versus cancerous tissue (Pub I), in good versus poor prognosis patients (Pub II), in tumours before and after cancer treatment (Pub III), and in frontal cortex versus hippocampus areas of the mouse brain (Pub IV).

In Pub I, 408 miRNAs were identified to have higher expression in cancerous versus normal breast tissue. Putative target mRNA with anti-correlated expression and with a minimum absolute log2 fold-change of 0.5 were identified from the target prediction database for a total of 4,208 pairings between 174 DE miRNAs and 915 transcripts. Most prominent were members of the miR-17-92 cluster of oncogenes (miR-17, miR-18a, miR-19a, miR-20a, miR-19b and miR-92a) and the cluster's human paralogs (specifically, miR-106a, miR106b, miR-93, miR-20b, miR-92a-2 and miR-363). Three commonly predicted targets of the miR-17-92 cluster showing anti-correlated expression were tumour suppressors *TGFBR2* [33], *DCN* [105], and *CAV1* [106] which were all under-expressed in the tumour tissues. This is consistent with previous studies presenting the miR-17-92 cluster as inhibitors of TGFβ signaling via regulation of key pathway genes [34] (Figure 6). Interestingly, while the role of *CAV1* in breast cancer is well-documented, it had not been previously linked to the miR-17-92 cluster.

In Pub III, three sets of interesting miRNAs were defined in the study: those DE between primary and relapsed tumours, those with relatively high expression in tumours compared to non-malignant B-cells (n = 24), and those with relatively low expression in tumours (n = 177). Differential analysis with both edgeR and

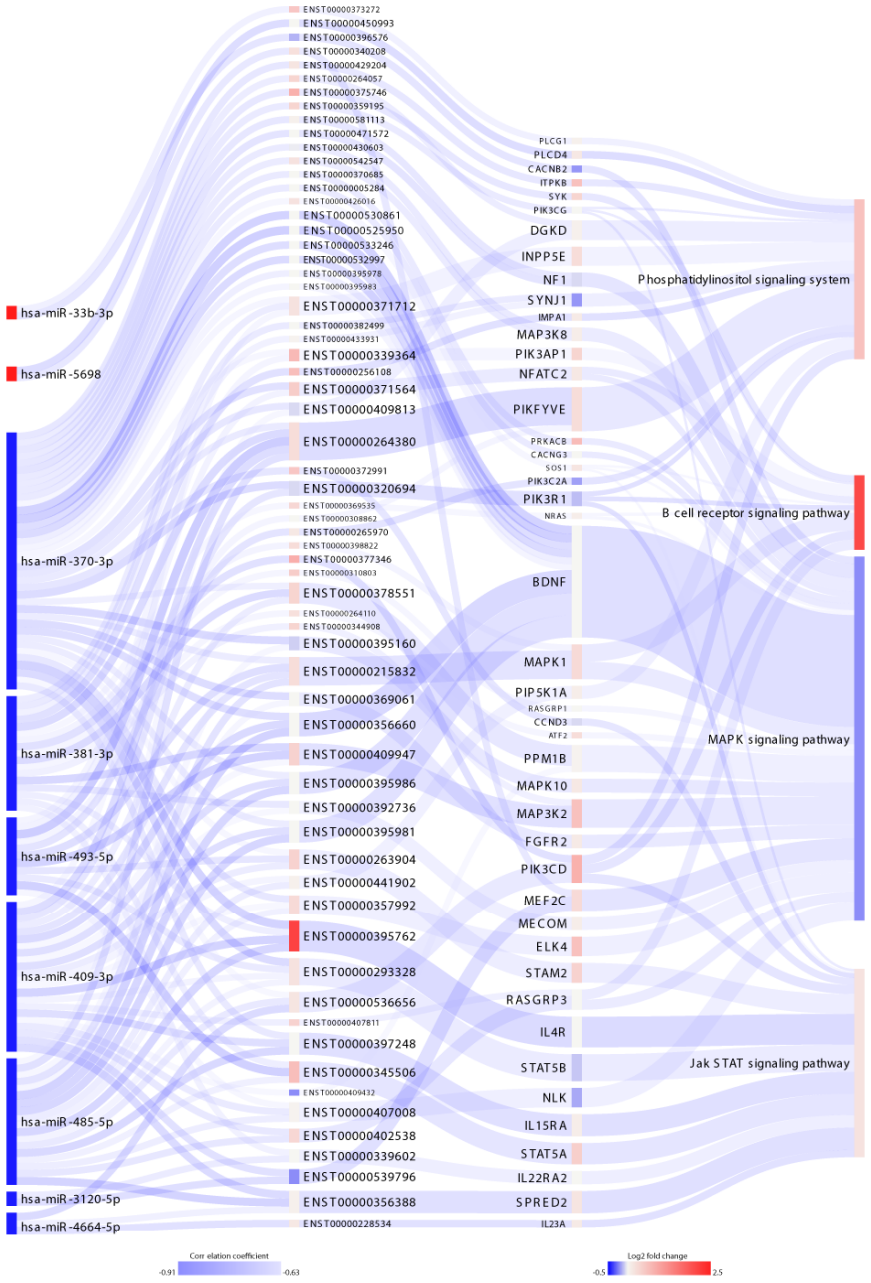

Figure 7: Combined DLBCL pathway of significant pathways in Pub III, annotated to targets of miRNAs differentially expressed in primary and relapsed tumours. Node colors are representative of gene, transcript, and pathway fold-change values with higher expression in primary tumours highlighted in blue and in relapsed tumours highlighted in red. MiRNAs have more extreme fold-change values but have been assigned the maximum and minimum colors such that red indicates a value $>>$ 2.5 and blue a value $<<$ -0.5.

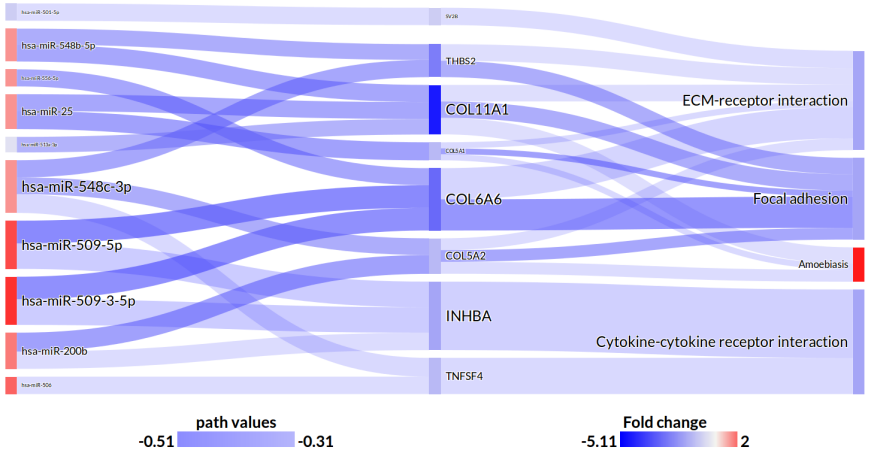

Figure 8: Genes differentially expressed between good and poor prognosis tumour samples in Pub II, their predicted targeting miRNAs with anti-correlated expression, and the pathways in which the genes are enriched in (false discovery rate < 10%). Node colours represent expression fold-change with higher expression in the good prognosis group highlighted in red and in the poor prognosis group in blue. Pathway nodes were given a fixed value of -2 if upregulated in poor prognosis samples. Paths between genes and pathways represent average expression correlation of targeting miRNAs. Paths connected to the focal adhesion pathway were selected to better identify associated miRNAs and predicted target genes.

DESeq identified 13 miRNAs with statistically significant expression change ($p < 0.05$). Among the high expression miRNAs were members of the oncogenic miR-10 family [38]. Among the low expression miRNAs were miR-129-5p [107], miR-663 [36], and miR-203a [108] with known tumour-suppressive roles.

Pathway impact analysis of the putative targets for both high and low expression miRNAs further supported oncogenic and tumour-suppressive roles, respectively. High expression miRNAs had putative targets enriched in cell adhesion while low expression miRNAs had putative targets enriched in cancer-associated MAPK signaling, cell cycle, and apoptosis pathways. MiRNAs DE between primary and relapsed tumours had putative targets enriched in lymphoma-associated pathways [109]: Phosphatidylinositol signaling, B-cell receptor signaling, and MAPK signaling (Figure 7). Validation by qRT-PCR confirmed lower expression for miR-409-3p, miR-381-3p, and miR-370-3p in relapsed DLBCL. Further functional analysis showed overexpression of the three validated miRNAs in DLBCL cells enhanced cell chemosensitivity.

The underlying aim of Pub IV was to see if NGS was a suitable platform for functional miRNA studies of different mouse brain tissue. The answer was yes, and covered the same benefits over microarray technology as described in Chapter

3.2. Several miRNA families and clusters were found to be DE in the mouse brain tissues and included the miR-8 family, miR-182-96-183 cluster, miR-212-312 cluster, and miR-34 family. Interestingly, more members from these families and clusters were found to be DE in small RNA-seq compared to microarray. Pathway analysis of the predicted targets of DE miRNAs for each brain region suggested regulation of brain region-specific signaling pathways [110, 111].

7.2.1 Director enables data visualization of regulatory cascades and guides further functional investigation

Co-regulation by miRNAs is a biological feature that is simpler to understand visually than statistically. Network graphs are generally used to represent gene regulation by miRNAs, but Sankey diagrams offer a more intuitive depiction of regulatory cascades with clear biological levels represented.

In Pub I, anti-correlated and predicted miRNA-target gene pairs of the TGFβ signaling pathway and mir-17/92 cluster were visualized with Director and revealed interesting differences in the co-regulation of tumour suppressors *TGFBR2* and *DCN* (Figure 6): 9 of the 10 expressed members appeared to target both of the known transcripts of *TGFBR2*, while all 10 miRNAs appeared to regulate just one of the four different transcripts of *DCN*.

In Pub II, 385 genes were identified to be DE between good and poor prognosis patients (false discovery rate $\leq 10\%$). Four KEGG pathways were found to be significantly enriched by 32 DE genes (false discovery rate $\leq 10\%$): Extracellular matrix-receptor interaction, Focal adhesion, Cytokine-cytokine receptor interaction, and Amoebiasis. Ten miRNAs with significant anti-correlated microarray expression ($p \leq 0.01$) were then identified for the 32 genes (Figure 8). While elevated levels of *COL11A1, INHBA*, and *THBS2* are a signature feature of metastasis [112], with collagen genes known to contribute to poor overall survival in OVCA [113], the putative connection to miRNAs had not yet been established. Data visualization with Director not only supported the idea of poor prognosis OVCA modulated through cell-matrix interaction pathways, but visually prioritized the identification of the above-mentioned oncogenes through the value-based color assignment of nodes and paths to depict the estimated contribution of a gene's altered expression on biological pathways.

In Pub III, visualization of the DE miRNAs and their target genes with Director (Figure 7) guided and supported further functional study in additional DLBCL cells. It identified miR-409-3p, miR-381-3p, and miR-370-3p as strong candidates for functional study and their co-regulatory behavior in the suppression of lymphoma-associated pathways.

7.2.2 Novel miRNA discovery identifies other small RNA with miRNA-like function

Novel miRNA discovery is configured in SePIA (Pub I) to first identify putative miRNA sequences for each sample and then join them across samples into overlapping genomic regions. The idea was to broaden novel miRNA discovery to include genomic regions that may contain other small RNAs either encoding miRNAs or containing miRNA-like features. Pub I identified 33 such novel miRNA regions DE between normal and tumour breast tissue (false-discovery rate < 0.05). Manual inspection of these regions revealed five overlapping small nucloeolar RNAs (*SNORA56, SNORA69, SNORD95, SNORD49, SNORD82*), two overlapping regions formerly annotated to putative miRNAs in Ensembl (*AL161626.1, FP236383.10*), and one overlapping small zinc-finger protein 813 (*ZNF813*).

Novel miRNA discovery was also performed but not included in the results of Pub III. One of the 10 putative novel miRNA regions DE between primary and relapse tumours annotated to *SNORD71*, a snoRNA with known miRNA-like features and a predicted role in methylation [114]. These results concur with previous work by Ender *et al.* [115] that identified a class of human snoRNAs with the ability to also function like miRNAs, and by Scott & Ono [116] that further defined a subset of snoRNAs as dual-function regulatory RNAs.

8 Discussion

The main aim of this work was to develop and use computational tools to better facilitate miRNA research, with emphasis on identifying potential miRNA biomarkers and therapeutic targets in two biologically complex datasets (mouse brain and human tumours). In doing so, the intent was not to recommend specific tools for a miRNA study, but rather to be inclusive of the *best representatives* — that is, reliable and established — of methods in use. In a field where technology and methods are still developing, and where biological features are still being discovered, it is more important in the long run to have a well-developed workflow that can robustly incorporate innovations in the field.

This thesis presents two tools to better facilitate miRNA research with multidimensional data, currently and in the future: SePIA, a workflow to implement standard processing, analysis, and integration of RNA and small RNA-seq data; and Director, an R package to visually explore the flow of biological information and provide context to the therapeutic potential of miRNAs.

SePIA has been applied in a variety of experimental settings. As our understanding of miRNA target recognition grows, a robust workflow such as SePIA becomes essential to accommodating the new technology and the consequent changes to computational algorithms resulting from them.

8.1 Implementation of standard RNA-seq procedures

Reproducibility is essential to any study based on RNA-seq data and, in the absence of a standard data processing pipeline, a standard procedure for implementing strategies and analysis scenarios is ideal. This work shows how modules with customizable component parts is a feasible and practical solution. By providing a stable infrastructure for the processing and analysis of RNA-seq data, a comparable level of output is achievable for similar RNA-seq projects.

Computational workflows are essentially a series of executable software, and SePIA was developed such that all intermediary results matched software expectations. That is, the files produced within the workflow could be replicated by executing the corresponding software independently with the same parameters. This is straightforward to do as parameter settings and software versions used are documented for each SePIA workflow instance.

While computational workflows have been previously done with RNA-seq [117, 118], small RNA-seq [119], and even integration of the two [120], the heavy computational requirements of combining all three is a major challenge to prospective

users. Some get around this by limiting dataset-specific customization to optimize usability and accessibility [57, 121]. However, with such variable experimental designs and sample sizes in a given project, this was not an acceptable limitation in the design of SePIA.

The incorporation of Docker [104] provided a simple solution to the dilemma of installing all the required tools with minimal computational proficiency. The Anduril workflow engine also provided a means to develop fully-automated reporting and documentation of the parameters executed in SePIA pipelines. These workflow conveniences help users to focus more on data analysis and hypothesis testing rather than computer specifications.

8.1.1 Workflow considerations and limitations

The choice of readily available software in the SePIA workflow is based on previously published comparisons of tools [45, 54, 122]. For example, Trimmomatic [91] works well with preprocessing both paired and single-end RNA-seq reads but was less efficient at trimming small RNA-seq reads to the expected miRNA length of 19-22nt compared to FASTX-Toolkit [90]. This is likely due to different stringencies in the detection of partial 3' adaptor sequences [122]. Modularity in the workflow allows for the implementation of novel methods not readily available in SePIA, or the implementation of similar tools in parallel, to process and analyze sequencing data.

Many variations of RNA-seq protocols and analyses have been published, making it challenging for new users to appreciate all of the steps necessary to conduct an RNA-seq study properly [50]. To serve as a base workflow template, SePIA pipelines are initially configured to use tools in general, case versus control, RNA-based expression studies.

8.2 Expression profiling miRNAs for clinically meaningful insights

This work set out to contribute to the area of miRNA research involved in the processing and analysis of expression data on a gene and transcript level. Context was placed on the biological function of miRNAs in complex and difficult-to-treat disease.

Aberrant miRNA expression is a recurring observation made in poor prognosis cancer types and those that enter metastasis [14, 69]. Analysis of expression profiles from different tissues and conditions make it possible to identify those molecules with aberrant expression potentially corresponding to important roles in

distinguishing tissue phenotype [8]. For example, differential analysis of miRNAs from Pub III and IV identified miRNAs with roles in cell growth and brain development, respectively. Biological function is well-annotated for protein-coding genes, so the inhibition of a predicted target's expression into a protein can be inferred as a miRNA's negative regulation of the gene's biological function. In this way, miRNAs have been identified in this work and others as viable biomarkers in disease diagnosis, treatment, and/or prognosis [14, 23, 25, 26].

While two approaches to differential expression are primarily used in this thesis work, the common assumption for all differential analyses is that samples being compared contain similar amounts of RNA. No single method dominates another across experimental designs, but edgeR and DESeq are consistently among the best performers [45, 48]. Additionally, both methods produced highly overlapping sets of differentially expressed miRNAs in this work.

The effect of a single miRNA on a single gene is sometimes negligible, especially in comparison to transcription factors, and has lead to speculation over the true biological contribution [123]. However, co-regulation of several miRNAs on several genes has been found to be significant [12, 32, 67] and is supported by the results of Pub II and III in cancer. Computational methods have been developed to predict such networks of miRNA-gene interactions from sequence and expression data [71, 124] but without the benefit of clear, one-to-one interactions that target identification methods provide, it can be a challenge to formulate hypotheses about the role of individual molecules and effectively prioritize functional investigation.

High-throughput computational approaches achieve the identification of many statistically significant driver molecules, but functional studies are still essential to prioritizing the list to relevant candidates [63]. The ability to conceptualize biological relevance from computational results is still an exclusively human skill [13, 43] . Director was developed on the idea that visualization provides a straightforward approach to integrating and interpreting result data. Work in Pub II and III demonstrate how a visualization tool combines expression analysis results into meaningful interpretations of the regulatory contributions of miRNAs on cancer pathways.

Complex diseases tend to operate on multiple molecular levels and so genes and transcripts may reveal only part of the story. As high-throughput technology becomes more affordable, large-scale multidimensional studies become more common. While these will add depth to our understanding of diseases, integration of such multidimensional data will also add new computational challenges. Both SePIA and Director were developed to help take on these challenges.

Recent evidence indicate that the abundance of proteins that make up the RISC

complex, including the miRNA-binding Argonaute proteins, has influence on the effectiveness of miRNA target-binding. This implies that some computationally predicted targets may indeed be true targets, but will not be detectable in cells where there is insufficient levels of RISC molecules [31]. To be detectable, functional regulation also seems to require target genes to have sufficient dose-sensitivity [123]. To my knowledge, protein data and tissue specificity have not yet been incorporated to target prediction algorithms, and no high-throughput method currently exists to identify dose-sensitive genes [123]. However, SePIA's robust and modular design guarantees it will be relatively easy to incorporate such data and tools (such as Director) to better characterize the regulatory role of miRNAs in the future. For example, results produced from a novel prediction algorithm could be incorporated as an additional resource for the miRNA-target gene database referenced in the miRNA-mRNA integration step. Predicted values for target binding strength could also be used as input for the paths connecting miRNAs to mRNAs in the regulatory cascades visualized with Director.

9 Conclusions

Scientific research is about furthering our knowledge through systematic collection and investigation. Data analysis and exploration leads to new (testable) hypotheses; new biological discoveries influence which and how measurement technologies and computational approaches are developed. Sometimes what we learn is that there are gaps in our knowledge, and to fill the gaps we must ask different questions [125]. The history of miRNA studies is a good example of this – though we are *more confident* with detecting small non-coding RNAs and functional roles in disease and development, we are *less certain* about why and how these functions are activated in different tissue context. MiRNA research has thus far only scratched the surface of an unexpectedly deep and complex regulatory system of gene expression [126]. The existence of other small, non-coding RNAs with miRNA-like features [115, 116, 127] adds to the mystique of miRNAs and further hints to a highly intricate and biologically relevant role for miRNAs.

Tremendous progress has been made in the last decade to improve the reputation of non-protein coding RNA. However, a lot of work is still needed to ascertain their exact biological functions and, consequently, clinical potential. The bottleneck facing biomedical research today is not the processing of large-scale data nor the availability of computational resources, but the mostly-manual and tedious interrogation of computational results for meaningful and actionable insights.

As the future of cancer research is directed more and more to personalized medicine, consequent development in both RNA-seq technologies and computational methods will bring us closer to achieving direct implications in clinical decision-making. With mounting evidence of the clinical potential of miRNAs and the first clinical trials already underway [26], their use in therapies may be realized sooner than our ability to precisely understand their regulatory behavior. The latter will still be necessary to optimize the use of miRNAs as biomarkers and treatment targets. Thus, contributions from this work will continue to facilitate a better understanding of the functional potential of miRNAs in novel, robust therapeutic strategies in all fields of biomedicine.

Acknowledgements

This work would not have been possible without the funding of the Doctoral Program in Biomedicine, The Academy of Finland (Center of Excellence in Cancer Genetics Research), the Sigrid Jusélius Foundation and Finnish Cancer Association.

The majority of this work was carried out in the Systems Biology Laboratory at the Faculty of Medicine, University of Helsinki. A heartfelt thank you to Professor Sampsa Hautaniemi for his inspiration and guidance. Your mentorship helped me grow confident in my abilities and independent in my work. This thesis is complete thanks to you.

Iiris Hovatta's lab gave me my first research position and started me on this doctoral path. I am grateful to Tero Aittokallio and Mikko Frilander for seeing my thesis through. I am especially grateful to Jesus Lopez, my first real mentor, who helped me overcome self-doubt and persevere.

I would like to thank my collaborators, Sirpa Leppä and her lab, for their huge efforts in the DLBCL project. Especially Suvi-Katri Leivonen, who was a pleasure to work with and I am honored to have as my co-author.

I have been fortunate to know many wonderful people in Biomedicum I. From the first day of Orientation, Virginia has been a good friend and a sympathetic ear. Past and present members of the Systems Biology Lab – Anna-Maria, Marko, Sirkku, Kristian, Riku, Javier, Lauri, Lilli, Julia, Chiara, Mikko, Emilia, Kayang, and Tiia – have been part of some of the best times I've had in the university. It is a special privilege to have published work with Ping, Alejandra, Ville, Amjad, and Chengyu. Thank you also to Rainer for always reminding us of biological relevance. The trips to and from coding camp were as exhilarating as they were insightful.

Finally, I am most thankful to those nearest and dearest to my heart. To my parents, Wilfredo and Isabelilia, thank you for nurturing a healthy desire for academic excellence, for encouraging travel and the pursuit of opportunities – even when it lead me to what might as well have been the farthest end of the world. To my awesome husband Sami, thank you for your love, support, and pokes of encouragement. None of this would have been possible without you. Last but certainly not least, thank you BB for setting us on our Next Big Adventure — because some things are just too exciting to wait.

Katherine Abigail Icay-Rouhianen
Helsinki, 2018

References

Page(s)

[1] Crick, F. H. (1958) On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138–163. 1

[2] Higgs, P. G & Lehman, N. (2015) The RNA World: molecular cooperation at the origins of life. *Nat. Rev. Genet.* **16**, 7–17. 1, 5

[3] Morris, K. V & Mattick, J. S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423–437. 1, 2, 5, 6, 10

[4] Mills, D. R, Peterson, R. L, & Spiegelman, S. (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. U.S.A.* **58**, 217–224. 1

[5] Mattick, J. S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2**, 986–991. 1, 3, 5

[6] Weinberg, R. A. (2014) *The Biology of Cancer*. (Garland Science), 2nd edition. 1, 2, 3

[7] The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. 1, 3

[8] Finotello, F & Di Camillo, B. (2015) Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* **14**, 130–142. 3, 11, 12, 13, 14, 26, 37

[9] International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. 3

[10] Venter, J. C, Adams, M. D, Myers, E. W, Li, P. W, Mural, R. J, Sutton, G. G, Smith, H. O, Yandell, M, Evans, C. A, Holt, R. A, Gocayne, J. D, Amanatides, P, Ballew, R. M, Huson, D. H, Wortman, J. R, Zhang, Q, Kodira, C. D, Zheng, X. H, Chen, L, Skupski, M, Subramanian, G, Thomas, P. D, Zhang, J, Gabor Miklos, G. L, Nelson, C, Broder, S, Clark, A. G, Nadeau, J, McKusick, V. A, Zinder, N, Levine, A. J, Roberts, R. J, Simon, M, Slayman, C, Hunkapiller, M, Bolanos, R, Delcher, A, Dew, I, Fasulo, D, Flanigan, M, Florea, L, Halpern, A, Hannenhalli, S, Kravitz, S, Levy, S, Mobarry, C, Reinert, K, Remington, K, Abu-Threideh, J, Beasley, E, Biddick, K, Bonazzi, V, Brandon, R, Cargill, M, Chandramouliswaran, I, Charlab, R, Chaturvedi, K, Deng, Z, Di Francesco, V, Dunn, P, Eilbeck, K, Evangelista, C, Gabrielian, A. E, Gan, W, Ge, W, Gong, F, Gu, Z, Guan, P, Heiman, T. J, Higgins, M. E, Ji, R. R, Ke, Z, Ketchum, K. A, Lai, Z, Lei, Y, Li, Z, Li, J, Liang, Y, Lin, X, Lu, F, Merkulov, G. V, Milshina, N, Moore, H. M, Naik, A. K, Narayan, V. A, Neelam, B, Nusskern, D, Rusch, D. B, Salzberg, S, Shao, W, Shue, B, Sun, J, Wang, Z, Wang, A, Wang, X, Wang, J, Wei, M, Wides, R, Xiao, C, Yan, C, Yao, A, Ye, J, Zhan, M, Zhang, W, Zhang, H, Zhao, Q, Zheng, L, Zhong, F, Zhong, W, Zhu, S, Zhao, S, Gilbert, D, Baumhueter, S, Spier, G, Carter, C, Cravchik, A, Woodage, T, Ali, F, An, H, Awe, A, Baldwin, D, Baden, H, Barnstead, M, Barrow, I, Beeson, K, Busam, D, Carver, A, Center, A, Cheng, M. L, Curry, L, Danaher, S, Davenport, L, Desilets, R, Dietz, S, Dodson, K, Doup,

L, Ferriera, S, Garg, N, Gluecksmann, A, Hart, B, Haynes, J, Haynes, C, Heiner, C, Hladun, S, Hostin, D, Houck, J, Howland, T, Ibegwam, C, Johnson, J, Kalush, F, Kline, L, Koduru, S, Love, A, Mann, F, May, D, McCawley, S, McIntosh, T, McMullen, I, Moy, M, Moy, L, Murphy, B, Nelson, K, Pfannkoch, C, Pratts, E, Puri, V, Qureshi, H, Reardon, M, Rodriguez, R, Rogers, Y. H, Romblad, D, Ruhfel, B, Scott, R, Sitter, C, Smallwood, M, Stewart, E, Strong, R, Suh, E, Thomas, R, Tint, N. N, Tse, S, Vech, C, Wang, G, Wetter, J, Williams, S, Williams, M, Windsor, S, Winn-Deen, E, Wolfe, K, Zaveri, J, Zaveri, K, Abril, J. F, Guigo, R, Campbell, M. J, Sjolander, K. V, Karlak, B, Kejariwal, A, Mi, H, Lazareva, B, Hatton, T, Narechania, A, Diemer, K, Muruganujan, A, Guo, N, Sato, S, Bafna, V, Istrail, S, Lippert, R, Schwartz, R, Walenz, B, Yooseph, S, Allen, D, Basu, A, Baxendale, J, Blick, L, Caminha, M, Carnes-Stine, J, Caulk, P, Chiang, Y. H, Coyne, M, Dahlke, C, Mays, A, Dombroski, M, Donnelly, M, Ely, D, Esparham, S, Fosler, C, Gire, H, Glanowski, S, Glasser, K, Glodek, A, Gorokhov, M, Graham, K, Gropman, B, Harris, M, Heil, J, Henderson, S, Hoover, J, Jennings, D, Jordan, C, Jordan, J, Kasha, J, Kagan, L, Kraft, C, Levitsky, A, Lewis, M, Liu, X, Lopez, J, Ma, D, Majoros, W, McDaniel, J, Murphy, S, Newman, M, Nguyen, T, Nguyen, N, Nodell, M, Pan, S, Peck, J, Peterson, M, Rowe, W, Sanders, R, Scott, J, Simpson, M, Smith, T, Sprague, A, Stockwell, T, Turner, R, Venter, E, Wang, M, Wen, M, Wu, D, Wu, M, Xia, A, Zandieh, A, & Zhu, X. (2001) The sequence of the human genome. *Science* **291**, 1304–1351.

3

[11] Ezkurdia, I, Juan, D, Rodriguez, J. M, Frankish, A, Diekhans, M, Harrow, J, Vazquez, J, Valencia, A, & Tress, M. L. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878.

3

[12] Stefani, G & Slack, F. J. (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.* **9**, 219–230.

3, 6, 7, 38

[13] Knox, S. S. (2010) From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int.* **10**, 11.

3, 11, 38

[14] Mazan-Mamczarz, K & Gartenhaus, R. B. (2013) Role of microRNA deregulation in the pathogenesis of diffuse large B-cell lymphoma (DLBCL). *Leuk. Res.* **37**, 1420–1428.

3, 10, 37, 38

[15] The Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120.

5

[16] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.

5

[17] Almeida, M. I, Reis, R. M, & Calin, G. A. (2011) MicroRNA history: discovery, recent applications, and next frontiers. *Mutat. Res.* **717**, 1–8.

5, 6, 9

[18] Lee, R. C, Feinbaum, R. L, & Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854.

5

[19] Wightman, B, Ha, I, & Ruvkun, G. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855–862.

5, 16

[20] Reinhart, B. J, Slack, F. J, Basson, M, Pasquinelli, A. E, Bettinger, J. C, Rougvie, A. E,

Horvitz, H. R, & Ruvkun, G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901–906. 5

[21] Pasquinelli, A. E, Reinhart, B. J, Slack, F, Martindale, M. Q, Kuroda, M. I, Maller, B, Hayward, D. C, Ball, E. E, Degnan, B, Muller, P, Spring, J, Srinivasan, A, Fishman, M, Finnerty, J, Corbo, J, Levine, M, Leahy, P, Davidson, E, & Ruvkun, G. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–89. 6

[22] Lai, E. C. (2002) Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30**, 363–364. 6, 10

[23] Schmitz, U, Woklenhauer, O, & Vera, J. (2013) *MicroRNA Cancer Regulation*, Advances in Experimental Medicine and Biology. (Springer Science + Business Media) Vol. 774. 6, 7, 10, 11, 16, 38

[24] Meltzer, P. S. (2005) Cancer genomics: small RNAs with big impacts. *Nature* **435**, 745–746. 6, 9, 10

[25] Diniz, G. P & Wang, D. Z. (2016) Regulation of Skeletal Muscle by microRNAs. *Compr Physiol* **6**, 1279–1294. 6, 10, 38

[26] Li, Z & Rana, T. M. (2014) Therapeutic targeting of microRNAs: current status and future challenges. *Nat Rev Drug Discov* **13**, 622–638. 6, 7, 10, 38, 40

[27] Bartel, D. P. (2004) Micrornas: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297. 7

[28] He, L, Sedwick, C, & He, L. (2015) Lin He: "Junk" DNA isn't. *J. Cell Biol.* **211**, 4–5. 7, 10, 18

[29] Hammond, S. M. (2015) An overview of microRNAs. *Adv. Drug Deliv. Rev.* **87**, 3–14. 7, 16, 17

[30] Bartel, D. P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233. 9, 17

[31] Ritchie, W & Rasko, J. E. (2014) Refining microRNA target predictions: sorting the wheat from the chaff. *Biochem. Biophys. Res. Commun.* **445**, 780–784. 10, 17, 18, 23, 39

[32] Li, X, Nie, J, Mei, Q, & Han, W. D. (2016) MicroRNAs: Novel immunotherapeutic targets in colorectal carcinoma. *World J. Gastroenterol.* **22**, 5317–5331. 10, 38

[33] Mogilyansky, E & Rigoutsos, I. (2013) The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death Differ.* **20**, 1603–1614. 10, 11, 31

[34] Dews, M, Fox, J. L, Hultine, S, Sundaram, P, Wang, W, Liu, Y. Y, Furth, E, Enders, G. H, El-Deiry, W, Schelter, J. M, Cleary, M. A, & Thomas-Tikhonenko, A. (2010) The myc-miR-17 92 axis blunts TGFbeta signaling and production of multiple TGFbeta-dependent antiangiogenic factors. *Cancer Res.* **70**, 8233–8246. 10, 11, 31

[35] Fuziwara, C. S & Kimura, E. T. (2015) Insights into Regulation of the miR-17-92 Cluster of miRNAs in Cancer. *Front Med (Lausanne)* **2**, 64. 10, 11

[36] Zang, W, Wang, Y, Wang, T, Du, Y, Chen, X, Li, M, & Zhao, G. (2015) miR-663 attenuates tumor growth and invasiveness by targeting eEF1A2 in pancreatic cancer. *Mol. Cancer* **14**,

10, 11, 33 37.

[37] Cao, H, Zhu, Q, Huang, J, Li, B, Zhang, S, Yao, W, & Zhang, Y. (2009) Regulation and functional role of eEF1A2 in pancreatic carcinoma. *Biochem. Biophys. Res. Commun.* **380**, 11 11–16.

[38] Tehler, D, HØyland-Kroghsbo, N. M, & Lund, A. H. (2011) The miR-10 microRNA 11, 33 precursor family. *RNA Biol* **8**, 728–734.

[39] Tsai, W. C, Hsu, S. D, Hsu, C. S, Lai, T. C, Chen, S. J, Shen, R, Huang, Y, Chen, H. C, Lee, C. H, Tsai, T. F, Hsu, M. T, Wu, J. C, Huang, H. D, Shiao, M. S, Hsiao, M, & Tsou, A. P. (2012) MicroRNA-122 plays a critical role in liver homeostasis and hepatocarcinogenesis. 11 *J. Clin. Invest.* **122**, 2884–2897.

[40] Johnson, S. M, Grosshans, H, Shingara, J, Byrom, M, Jarvis, R, Cheng, A, Labourier, E, Reinert, K. L, Brown, D, & Slack, F. J. (2005) RAS is regulated by the let-7 microRNA 11 family. *Cell* **120**, 635–647.

[41] Hoheisel, J. D. (2006) Microarray technology: beyond transcript profiling and genotype 11, 12 analysis. *Nat. Rev. Genet.* **7**, 200–210.

12, 13, 14 [42] McPherson, J. D. (2009) Next-generation gap. *Nat Methods* **6**, 2–5.

[43] Goodwin, S, McPherson, J. D, & McCombie, W. R. (2016) Coming of age: ten years of 12, 38 next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351.

[44] Dillies, M. A, Rau, A, Aubert, J, Hennequet-Antier, C, Jeanmougin, M, Servant, N, Keime, C, Marot, G, Castel, D, Estelle, J, Guernec, G, Jagla, B, Jouneau, L, Laloe, D, Le Gall, C, Schaeffer, B, Le Crom, S, Guedj, M, & Jaffrezic, F. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics* **13, 14** **14**, 671–683.

[45] Soneson, C & Delorenzi, M. (2013) A comparison of methods for differential expression 13, 15, 26, 27, analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91. 37, 38

[46] Engstrom, P. G, Steijger, T, Sipos, B, Grant, G. R, Kahles, A, Ratsch, G, Goldman, N, Hubbard, T. J, Harrow, J, Guigo, R, Bertone, P, Alioto, T, Behr, J, Bertone, P, Bohnert, R, Campagna, D, Davis, C. A, Dobin, A, Engstrom, P. G, Gingeras, T. R, Goldman, N, Grant, G. R, Guigo, R, Harrow, J, Hubbard, T. J, Jean, G, Kahles, A, Kosarev, P, Li, S, Liu, J, Mason, C. E, Molodtsov, V, Ning, Z, Ponstingl, H, Prins, J. F, Ratsch, G, Ribeca, P, Seledtsov, I, Sipos, B, Solovyev, V, Steijger, T, Valle, G, Vitulo, N, Wang, K, Wu, T. D, & Zeller, G. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. 13 *Nat. Methods* **10**, 1185–1191.

[47] Helwak, A & Tollervey, D. (2014) Mapping the miRNA interactome by cross-linking 14, 17 ligation and sequencing of hybrids (CLASH). *Nat Protoc* **9**, 711–728.

[48] Rapaport, F, Khanin, R, Liang, Y, Pirun, M, Krek, A, Zumbo, P, Mason, C. E, Socci, N. D, & Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis 14, 26, 27, 38 methods for RNA-seq data. *Genome Biol.* **14**, R95.

[49] Rantanen, V, Valori, M, & Hautaniemi, S. (2014) Anima: modular workflow system for comprehensive image data analysis. *Front Bioeng Biotechnol* **2**, 25. 14

[50] Conesa, A, Madrigal, P, Tarazona, S, Gomez-Cabrero, D, Cervera, A, McPherson, A, Szczesniak, M. W, Gaffney, D. J, Elo, L. L, Zhang, X, & Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13. 15, 27, 37

[51] Milicchio, F, Rose, R, Bian, J, Min, J, & Prosperi, M. (2016) Visual programming for next-generation sequencing data analytics. *BioData Min* **9**, 16. 15, 30

[52] Ovaska, K, Laakso, M, Haapa-Paananen, S, Louhimo, R, Chen, P, Aittomaki, V, Valo, E, Nunez-Fontarnau, J, Rantanen, V, Karinen, S, Nousiainen, K, Lahesmaa-Korpinen, A. M, Miettinen, M, Saarinen, L, Kohonen, P, Wu, J, Westermarck, J, & Hautaniemi, S. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med* **2**, 65. 15

[53] Tam, S, Tsao, M. S, & McPherson, J. D. (2015) Optimization of miRNA-seq data preprocessing. *Brief. Bioinformatics* **16**, 950–963. 15

[54] Ye, H, Meehan, J, Tong, W, & Hong, H. (2015) Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. *Pharmaceutics* **7**, 523–541. 15, 37

[55] Lee, R. C & Ambros, V. (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862–864. 16

[56] Hofacker, I. L, Fontana, W, Stadler, P. F, Bonhoeffer, L. S, Tacker, M, & Schuster, P. (1994) Fast folding and comparison of rna secondary structures. *Monatsh Chem* **125**, 167–188. 16

[57] Hackenberg, M, Rodriguez-Ezpeleta, N, & Aransay, A. M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.* **39**, W132–138. 16, 28, 37

[58] Lewis, B. P, Burge, C. B, & Bartel, D. P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20. 17

[59] Grimson, A, Farh, K. K, Johnston, W. K, Garrett-Engele, P, Lim, L. P, & Bartel, D. P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**, 91–105. 17, 24

[60] John, B, Enright, A. J, Aravin, A, Tuschl, T, Sander, C, & Marks, D. S. (2004) Human MicroRNA targets. *PLoS Biol.* **2**, e363. 18, 24

[61] Kertesz, M, Iovino, N, Unnerstall, U, Gaul, U, & Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nat Genet.* **39**, 1278–1284. 18, 24

[62] Hsu, S. D, Tseng, Y. T, Shrestha, S, Lin, Y. L, Khaleel, A, Chou, C. H, Chu, C. F, Huang, H. Y, Lin, C. M, Ho, S. Y, Jian, T. Y, Lin, F. M, Chang, T. H, Weng, S. L, Liao, K. W, Liao, I. E, Liu, C. C, & Huang, H. D. (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* **42**, 78–85. 18, 24

19, 38 [63] Ledford, H. (2010) Big science: The cancer genome challenge. *Nature* **464**, 972–974.

19 [64] Cairo, A. (2016) *The Truthful Art: Data, Charts, and Maps for Communication*. (New Riders Publishing), 1 edition.

19 [65] Chelaru, F, Smith, L, Goldstein, N, & Bravo, H. C. (2014) Epiviz: interactive visual analytics for functional genomics data. *Nat. Methods* **11**, 938–940.

19 [66] Bostock, M, Ogievetsky, V, & Heer, J. (2011) D3: Data-Driven Documents. *IEEE Trans Vis Comput Graph* **17**, 2301–2309.

22, 38 [67] The Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615.

23, 25 [68] Kozomara, A & Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, 68–73.

23, 37 [69] Yang, D, Sun, Y, Hu, L, Zheng, H, Ji, P, Pecot, C. V, Zhao, Y, Reynolds, S, Cheng, H, Rupaimoole, R, Cogdell, D, Nykter, M, Broaddus, R, Rodriguez-Aguayo, C, Lopez-Berestein, G, Liu, J, Shmulevich, I, Sood, A. K, Chen, K, & Zhang, W. (2013) Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell* **23**, 186–199.

23, 28 [70] Tusher, V. G, Tibshirani, R, & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5116–5121.

23, 38 [71] Muniategui, A, Pey, J, Planes, F. J, & Rubio, A. (2013) Joint analysis of miRNA and mRNA expression data. *Brief Bioinformatics* **14**, 263–278.

23 [72] Wang, Y. P & Li, K. B. (2009) Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics* **10**, 218.

23 [73] Cascione, L, Gasparini, P, Lovat, F, Carasi, S, Pulvirenti, A, Ferro, A, Alder, H, He, G, Vecchione, A, Croce, C. M, Shapiro, C. L, & Huebner, K. (2013) Integrated microRNA and mRNA signatures associated with survival in triple negative breast cancer. *PLoS ONE* **8**, e55910.

24 [74] Miranda, K. C, Huynh, T, Tay, Y, Ang, Y. S, Tam, W. L, Thomson, A. M, Lim, B, & Rigoutsos, I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217.

24 [75] Kruger, J & Rehmsmeier, M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* **34**, W451–454.

24 [76] Wang, X. (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**, 1012–1017.

24 [77] Maragkakis, M, Reczko, M, Simossis, V. A, Alexiou, P, Papadopoulos, G. L, Dalamagas, T, Giannopoulos, G, Goumas, G, Koukis, E, Kourtis, K, Vergoulis, T, Koziris, N, Sellis, T, Tsanakas, P, & Hatzigeorgiou, A. G. (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.* **37**, W273–276.

[78] Peterson, S. M, Thompson, J. A, Ufkin, M. L, Sathyanarayana, P, Liaw, L, & Congdon,

C. B. (2014) Common features of microRNA target prediction tools. *Front Genet* **5**, 23. 24

[79] Dobin, A, Davis, C. A, Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S, Batut, P, Chaisson, M, & Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21. 24, 28

[80] Langmead, B, Trapnell, C, Pop, M, & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25. 24, 28

[81] Anders, S, Pyl, P. T, & Huber, W. (2014) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 25, 28

[82] Robinson, M. D, McCarthy, D. J, & Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. 25, 27, 28

[83] Anders, S & Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106. 25, 27, 28

[84] Bullard, J. H, Purdom, E, Hansen, K. D, & Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94. 25

[85] Gautier, L, Cope, L, Bolstad, B. M, & Irizarry, R. A. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315. 26, 28

[86] Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3. 26, 28

[87] Oshlack, A, Robinson, M. D, & Young, M. D. (2010) From RNA-seq reads to differential expression results. *Genome Biol.* **11**, 220. 27

[88] Love, M. I, Huber, W, & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. 27, 28

[89] Zhou, X, Lindsay, H, & Robinson, M. D. (2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42**, e91. 27

[90] Hannon Lab. (2009) FASTX Toolkit: FASTQ/A short-reads pre-processing tools (`http://hannonlab.cshl.edu/fastx_toolkit/`). 28, 37

[91] Bolger, A. M, Lohse, M, & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. 28, 37

[92] Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc`). 28

[93] Trapnell, C, Pachter, L, & Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111. 28

[94] Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G, & Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. 28

[95] Broad Institute. (2009) Picard Toolkit (`http://broadinstitute.github.io/`

28 picard/).

[96] DeLuca, D. S, Levin, J. Z, Sivachenko, A, Fennell, T, Nazaire, M. D, Williams, C, Reich, M, Winckler, W, & Getz, G. (2012) RNA-SeQC: RNA-seq metrics for quality control and
28 process optimization. *Bioinformatics* **28**, 1530–1532.

[97] Wang, L, Wang, S, & Li, W. (2012) RSeQC: quality control of RNA-seq experiments.
28 *Bioinformatics* **28**, 2184–2185.

[98] Trapnell, C, Williams, B. A, Pertea, G, Mortazavi, A, Kwan, G, van Baren, M. J, Salzberg, S. L, Wold, B. J, & Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat.*
28 *Biotechnol.* **28**, 511–515.

[99] Edmonson, M. N, Zhang, J, Yan, C, Finney, R. P, Meerzaman, D. M, & Buetow, K. H. (2011) Bambino: a variant detector and alignment viewer for next-generation sequencing
28 data in the SAM/BAM format. *Bioinformatics* **27**, 865–866.

[100] Trapnell, C, Hendrickson, D. G, Sauvageau, M, Goff, L, Rinn, J. L, & Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat*
28 *Biotechnol.* **31**, 46–53.

[101] Grothendieck, G. (2014) *sqldf: Perform SQL selects on R Data Frames*. R package version
28 0.4-7.1.

[102] Tarca, A. L, Draghici, S, Khatri, P, Hassan, S. S, Mittal, P, Kim, J. S, Kim, C. J, Kusanovic, J. P, & Romero, R. (2009) A novel signaling pathway impact analysis. *Bioinformatics* **25**,
28 75–82.

[103] Qiagen Bioinformatics. (2010) Ingenuity Pathway Analysis (https://www.
28 qiagenbioinformatics.com/products/ingenuity-pathway-analysis/).

[104] Merkel, D. (2014) Docker: lightweight linux containers for consistent development and
30, 37 deployment. *Linux Journal* **2014**, 2.

[105] Grant, D. S, Yenisey, C, Rose, R. W, Tootell, M, Santra, M, & Iozzo, R. V. (2002) Decorin
31 suppresses tumor cell-mediated angiogenesis. *Oncogene* **21**, 4765–4777.

[106] Bai, L, Deng, X, Li, Q, Wang, M, An, W, Deli, A, Gao, Z, Xie, Y, Dai, Y, & Cong, Y. S. (2012) Down-regulation of the cavin family proteins in breast cancer. *J. Cell. Biochem.* **113**,
31 322–328.

[107] Hedstrom, G, Thunberg, U, Berglund, M, Simonsson, M, Amini, R. M, & Enblad, G. (2013) Low expression of microRNA-129-5p predicts poor clinical outcome in diffuse large B cell
33 lymphoma (DLBCL). *Int. J. Hematol.* **97**, 465–471.

[108] Diao, Y, Guo, X, Jiang, L, Wang, G, Zhang, C, Wan, J, Jin, Y, & Wu, Z. (2014) miR-203, a tumor suppressor frequently down-regulated by promoter hypermethylation in
33 rhabdomyosarcoma. *J. Biol. Chem.* **289**, 529–539.

[109] Rossi, D, Ciardullo, C, & Gaidano, G. (2013) Genetic aberrations of signaling pathways in lymphomagenesis: revelations from next generation sequencing studies. *Semin. Cancer*

Biol. **23**, 422–430. 33

[110] Xu, J & Wong, C. (2008) A computational screen for mouse signaling pathways targeted by microRNA clusters. *RNA* **14**, 1276–1283. 34

[111] Olsen, L, Klausen, M, Helboe, L, Nielsen, F. C, & Werge, T. (2009) MicroRNAs show mutually exclusive expression patterns in the brain of adult male rats. *PLoS ONE* **4**, e7225.34

[112] Kim, H, Watkinson, J, Varadan, V, & Anastassiou, D. (2010) Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med Genomics* **3**, 51. 34

[113] Cheon, D. J, Tong, Y, Sim, M. S, Dering, J, Berel, D, Cui, X, Lester, J, Beach, J. A, Tighiouart, M, Walts, A. E, Karlan, B. Y, & Orsulic, S. (2014) A collagen-remodeling gene signature regulated by TGF-β signaling is associated with metastasis and poor survival in serous ovarian cancer. *Clin. Cancer Res.* **20**, 711–723. 34

[114] Huttenhofer, A, Kiefmann, M, Meier-Ewert, S, O'Brien, J, Lehrach, H, Bachellerie, J. P, & Brosius, J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* **20**, 2943–2953. 35

[115] Ender, C, Krek, A, Friedlander, M. R, Beitzinger, M, Weinmann, L, Chen, W, Pfeffer, S, Rajewsky, N, & Meister, G. (2008) A human snoRNA with microRNA-like functions. *Mol. Cell* **32**, 519–528. 35, 40

[116] Scott, M. S & Ono, M. (2011) From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie* **93**, 1987–1992. 35, 40

[117] Goncalves, A, Tikhonov, A, Brazma, A, & Kapushesky, M. (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* **27**, 867–869. 36

[118] Zhao, W, Liu, W, Tian, D, Tang, B, Wang, Y, Yu, C, Li, R, Ling, Y, Wu, J, Song, S, & Hu, S. (2011) wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics* **27**, 3076–3077. 36

[119] Giurato, G, De Filippo, M. R, Rinaldi, A, Hashim, A, Nassa, G, Ravo, M, Rizzo, F, Tarallo, R, & Weisz, A. (2013) iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics* **14**, 362. 36

[120] Chae, H, Rhee, S, Nephew, K. P, & Kim, S. (2015) BioVLAB-MMIA-NGS: microRNA-mRNA integrated analysis using high-throughput sequencing data. *Bioinformatics* **31**, 265–267. 36

[121] Goecks, J, Nekrutenko, A, Taylor, J, Afgan, E, Ananda, G, Baker, D, Blankenberg, D, Chakrabarty, R, Coraor, N, Goecks, J, Von Kuster, G, Lazarus, R, Li, K, Nekrutenko, A, Taylor, J, & Vincent, K. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86. 37

[122] Lindgreen, S. (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes* **5**, 337. 37

[123] Pinzon, N, Li, B, Martinez, L, Sergeeva, A, Presumey, J, Apparailly, F, & Seitz, H. (2017) microRNA target prediction programs predict many false positives. *Genome Res.* **27**, 38, 39 234–245.

[124] Huang, J. C, Babak, T, Corson, T. W, Chua, G, Khan, S, Gallie, B. L, Hughes, T. R, Blencowe, B. J, Frey, B. J, & Morris, Q. D. (2007) Using expression profiling data to 38 identify human microRNA targets. *Nat. Methods* **4**, 1045–1049.

[125] Gleiser, M. (2015) *The Island of Knowledge: The Limits of Science and the Search for* 40 *Meaning*. (Basic Books), 1 edition.

[126] Steinkraus, B. R, Toegel, M, & Fulga, T. A. (2016) Tiny giants of gene regulation: experimental strategies for microRNA functional studies. *Wiley Interdiscip Rev Dev Biol* **5**, 40 311–362.

[127] Bratkovic, T & Rogelj, B. (2014) The many faces of small nucleolar RNAs. *Biochim.* 40 *Biophys. Acta* **1839**, 438–443.