Investigating small non-coding RNA regulating the hallmarks of cancer

Andrew Dhawan
St Hugh's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2018

This thesis is dedicated to
all those who inspire and help me to be
a better scientist and doctor.

Acknowledgements

I remember learning about cancer and the genetic changes that cause it in a high school biology class c.2007, taught by the inimitable Mrs. Quinn. Since that day, I was hooked. I was insatiably curious about why small changes in cells led to the massive consequence of cancer. I am grateful to you, Mrs. Quinn, for introducing me to this problem in a way that left me with an endless appetite for more knowledge. In the following years, I met many friends, mentors, and colleagues who helped me become the scientist I am today.

Dr. Brian Ingalls and Dr. Abdullah Hamadeh, thank you for giving me my first chance at research and teaching me the basics of mathematical biology, and how to be a researcher. Dr. Siv Sivaloganathan and Dr. Mohammad Kohandel, thank you for letting me spend so many summers working and learning the art of mathematical oncology from you. Dr. Colin Phipps and Dr. Venkata Manem, thank you for being such wonderful friends and supports along the way - every question I had, you worked tirelessly to help me answer. Dr. Yifan Li, you have been with me every step of the way on this adventure; thank you for always being there for me (and your bone-crushing hugs).

After my undergraduate studies, I started in medical school, where I had the immense privilege of learning the art of medicine from talented, caring clinicians and allied health professionals, and most of all, from the patients along the way. I am grateful for all those at Queen's who taught me the secrets of the body in health and disease. To the patients I have seen, my work is driven only by the inspiration and strength from you; thank you for letting me share in your journeys. I am also grateful to the 99 colleagues alongside whom I studied medicine; my best friends in the most turbulent of times - Dr. A. MacDonald, Dr. M. Mikhaeil, Dr. A. Khanna, and Dr. K. Shaikh, just to name a few.

Following medical school, I would not have been able to make it to Oxford without the generosity and kindness of those giving to the Clarendon fund and Cancer Research UK, allowing me the financial means to pursue this degree - thank you. Thank you to all of my friends in Oxford who have supported me in so many ways with kind words, helpful discussions over tea, and diversions from work. Truly, these are moments I will forever cherish; thank you Dr. J.P. Taylor-King, V. Ma, S. Ramachandran, A. Sekar, and so many others.

Thank you to my labmates and friends in both Oxford and Cleveland for keeping every day interesting, providing unending company through the long work hours of the day, and fun outside of work - especially you, N. Boros.

To my supervisors based in Oxford, Dr. Francesca Buffa and Dr. Adrian Harris, thank you for your intellectual discussions, and your openness to working with me. Truly, without you, none of this would have been possible. To Dr. Jake Scott, my supervisor based at the Cleveland Clinic, thank you for answering my first email to you c.2013 - I had no idea at the time, but that was one of the greatest and most life-changing things to happen to me. I cannot thank you enough. Your support has brought my career to new heights, and I'm so grateful to work with you every day.

Most of all, to my family, for their unending, infinite love and support – I am, and forever will be, thankful.

Thank you for this privilege.

Abstract

Non-coding RNA are increasingly being implicated as key regulators of the human transcriptome, and are thought to play important roles in the pathogenesis and progression of cancer. Understanding how non-coding RNA, specifically microRNA (miRNA) and circular RNA (circRNA), contribute to cancer is an area of significant clinical interest.

In this work, I utilise existing large genomic datasets, generated from the Cancer Genome Atlas and Metabric projects to gain insight into the associations between coding and non-coding RNA. In the first chapter, I develop a statistical methodology (*sigQC*), by which the applicability of mRNA gene expression signatures to datasets and the quality of this application can be assessed. In a second chapter, I use this methodology to generate a statistical mapping from miRNA to gene signatures for the hallmarks of cancer. I show that there is a core set of miRNA statistically associated with these gene signatures, which preferentially and, in some cases, exclusively, anti-correlate across cancer types with a subset of tumour suppressor genes, such as *PTEN*, *FAT4*, and *CDK12*.

In the next chapter, I build upon emerging knowledge of the changes to miRNA biogenesis in hypoxia, involving amplification of *AGO2* and codeletion of *DICER1*, and determine the miRNA statistically associated both with these copy number changes and with hypoxia gene expression. In doing so, I have statistically associated miRNA which have mature form associated with poor prognosis, increased invasiveness, and metastasis. Following this, I carry out an analysis of circRNA in breast cancers, uncovering statistical evidence which may be indicative of an autoregulatory feedback mechanism associated with hypoxia gene signature expression, involving a circRNA antisense to the *HSP90AB1* gene.

Having identified the potential involvement of a circRNA, a species thought to function as repressors, or sponges for miRNA molecules, I characterised the possible effects of generalised miRNA sponges on the dynamics of a

non-coding RNA (ncRNA) feedback loop. By analysing a system of delay differential equations in the deterministic and stochastic settings, I uncovered the range of potential dynamics that this feedback loop could exhibit. I showed that different kinetic properties for miRNA sponges gave rise to different behaviours of this network, including transient or sustained oscillations, which may have important roles in cancer and developmental biology.

Contents

List of Figures

List of Tables

Acronyms

AGO2 Argonaute-2.

CDF cumulative distribution function.

cDNA complementary DNA.

ceRNA competing endogenous mRNA.

circRNA circular RNA.

CNV copy number variant.

COSMIC catalogue of somatic mutations in cancer.

DICER1 DICER-1 RNA helicase.

DNA deoxyribonucleic acid.

EGA European Genome-Phenome Archive.

EMT epithelial-mesenchymal transition.

ER estrogen receptor.

FPKM fragments per kilobase million.

GDAC Broad Institute genomic data access centre.

GEO gene expression omnibus.

GSEA gene set enrichment analysis.

GSVA gene set variation analysis.

HSP90AB1 Heat shock protein 90 AB1.

ICGC International Cancer Genome Consortium.

LHS Latin hypercube sampling.

lncRNA long non-coding RNA.

miRNA microRNA.

mRNA messenger RNA.

MSigDB molecular signatures database.

ncRNA non-coding RNA.

NGS next-generation sequencing.

ODE ordinary differential equation.

PARP poly-adenosine diphosphate ribose polymerase.

PC principal components.

PCA principal components analysis.

PCA1 first principal component.

PCR polymerase chain reaction.

PDE partial differential equation.

PLAGE pathway level analysis of gene expression.

polyA polyadenylated.

PR progesterone receptor.

pre-miRNA precursor miRNA.

pri-miRNA primary miRNA.

RISC RNA-induced silencing complex.

RMA robust multiarray average.

RNA ribonucleic acid.

RNA-seq RNA sequencing.

RPM reads per million.

rRNA ribosomal RNA.

RSEM RNA-seq by expectation-maximisation.

siRNA small interfering RNA.

SNP single nucleotide polymorphism.

ssGSEA single sample gene set enrichment analysis.

TCGA The Cancer Genome Atlas.

TPM transcripts per million.

TSG tumour suppressor genes.

UTR untranslated region.

VST variance-stabilising transforms.

Chapter 1

Introduction

1.1 What is cancer?

Cancer is a disease that, at its core, is the result of uncontrolled proliferation of cells, and the spread of these aberrant cells to disseminated sites within the body. These rogue cells can result in organ dysfunction, immune response-related symptoms, increased metabolic demand, and ultimately, death. Decades of research into how normal cells transform into malignant cells has shown the root cause to be aberrations of the genomes of these cells; a faulty instruction set regulating their function.

In cancer, specific pathways related to proliferation, such as the growth factor receptor pathways (e.g. *EGFR* [1]), and pathways related to apoptosis (cellular death), such as *BCL-2*, *BAD*, and *BAX*, are dysregulated as a result of changes at the DNA level [2]. Small changes in the cell's instruction set have profound impacts on the activities of these proteins, and downstream effects of these changes lead cells to proliferate without bound. Ultimately, the proliferation of cancer cells changes their environment, causing selective forces acting on the growing tumour to drive these cells to invade and spread to distant sites throughout the body, often resulting in death of the host [3, 4].

1.1.1 The hallmarks of cancer summarise the potential behaviours of cancer cells

To distill the enormous complexity inherent across the many possible phenotypes of cancer, Hanahan and Weinberg described a set of phenotypic aberrations, which they termed the hallmarks of cancer [4, 5]. The initial set of hallmarks of cancer included 6 traits: sustaining proliferative signalling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, and

resisting cell death, as summarised in Figure 1.1A. An updated set in 2011 added four traits representative of the emerging understanding of cancer biology: avoiding immune destruction, tumour promoting inflammation, genome instability and mutation, and deregulating cellular energetics (Figure 1.1B). Together, these two landmark articles have provided researchers with the intrinsic understanding of the phenotypic behaviours of cancer cells. Subsequently, cancer research has focussed broadly upon identifying these phenotypes; how they arise, and how they may be targeted by drug and radiation therapy.

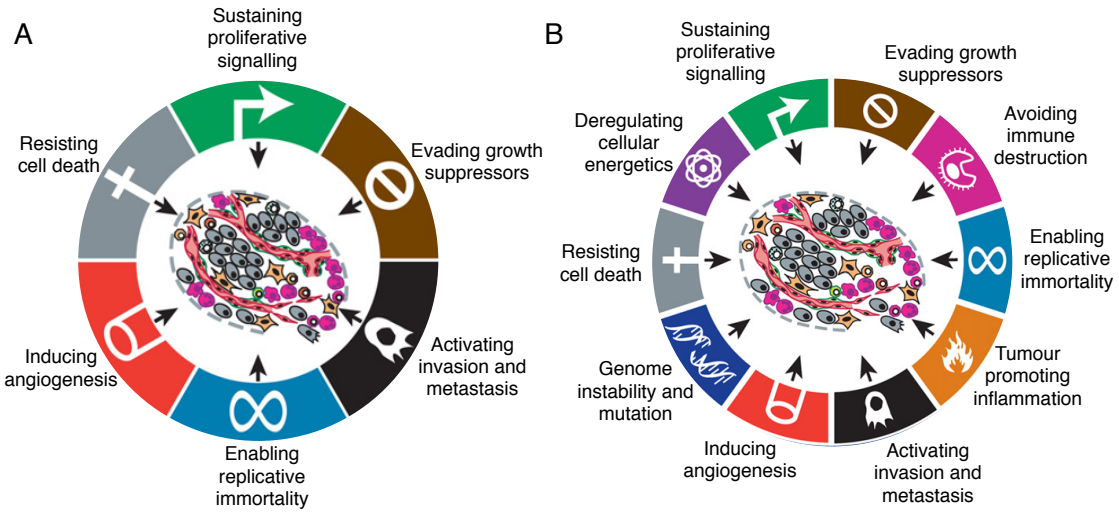

Figure 1.1: **The hallmarks of cancer, original edition and next generation, as per Hanahan and Weinberg.** Image adapted from Hanahan and Weinberg [5]. a) The original six hallmarks of cancer, published in 2000. b) The next generation hallmarks of cancer, published in the updated review in 2011. Figure reproduced with permission.

1.1.2 Criticism of the hallmarks of cancer

To present a balanced picture of the hallmarks of cancer, a summary of the criticisms from the broader cancer research community about using these hallmarks to characterise cancer cells must also be presented.

A well-known criticism of the hallmarks of cancer was written by Lazebnik as a commentary in response to the original work by Hanahan and Weinberg, and argued that the hallmarks identified are not unique to malignant solid tumours; and that therefore these cannot be hallmarks [6]. More specifically, Lazebnik argues that as a negative control for a hallmark, one should consider a benign solid tumour, and in this case, nearly every one of Hanahan and Weinberg's hallmarks are also features

of these benign lesions [6]. For instance, owing to the fact that benign lesions reach very large sizes, angiogenesis and the hypoxic response are features of these tumours, which also have a near limitless replicative potential [6]. Of the initial six hallmarks, Lazebnik argues that only invasion and metastasis pass this test of negative control; and notes that one must be very wary of interchanging the terms tumour and cancer, which represent two very different biological entities [6].

A second well-known criticism of the hallmarks of cancer was written by Sonnenschein and Soto in [7]. This criticism centres upon the idea that cancer is really a tissue-level disorder, and that characterising the disease process at the cellular level loses the nuance of tissue-level interactions and the evolutionary nature of the disease. They argue in favour of a model called the tissue organisation field theory, and state that instead of the traits that Hanahan and Weinberg attribute to cancers, such as proliferation and motility, that these are actually intrinsic to all cells, and cancer arises due to changes in the ways cells interact with each other in a tissue [7]. A more macroscopic, ecologically-defined, and non-cell autonomous view, they argue, is essential for understanding and characterising cancer.

1.2 Genetics of cancer

The central dogma of molecular biology states that from the instruction set defined by the DNA within a cell, RNA molecules are transcribed, which are ultimately translated into proteins, and these carry out a plethora of functions; effecting this instruction set [8]. Cancer has been shown to often arise as a result of the dysregulation of specific proteins, such as oncogenes and tumour suppressor genes; for instance *KRAS*, *TP53*, and *RET* [5]. The mutated forms of oncogenes are generally constitutively active forms of the molecules, leading to activation of downstream growth pathways. Likewise, the mutated form of tumour suppressor genes leads to their loss of function in preventing DNA damage or loss of cellular growth checkpoints. As a result, the last 50 years of research have focussed on developing tools to interrogate the expression of these proteins, such as functional assays for downstream members of these pathways and their mutational and copy number status [9]. In particular, research has shown that, at the level of the RNA coding for the protein, the factors leading to altered expression of the RNA molecule, such as DNA methylation, DNA binding proteins, RNA binding proteins, transcription factors, and changes to enhancer regions of the genome may all potentially contribute to carcinogenesis [5]. At the level of DNA, carcinogenesis is thought to occur because of mutations, large scale

genetic changes, such as copy number changes, or translocations, leading to altered RNA molecules and protein products [10, 11].

1.2.1 Cancer is a disease driven by evolution

Underscoring the genetic basis cancer is the evolutionary model of tumour progression [12, 13, 14]. That is, cancer is seen as a micro-scale model system for evolution and natural selection [12]. Because cancer cells are highly plastic and rapidly dividing, they constantly adapting and responding to dynamic microenvironmental conditions with rapid turnover, such that there is continuous natural selection occurring within the tumour [15]. This evolutionary process ultimately leaves its mark in the form of heritable changes in cells, typically through mutational changes in DNA [3, 16]. Recent studies have shown that this process can indeed be measured, reconstructed, and modelled computationally with potential to design improved disease biomarkers [14, 17, 18]. The ability to measure tumour evolution is currently being examined for its clinical utility, through studies of circulating tumour DNA and multiple spatially-distinct biopsies through projects led by the TracerX consortium, for instance [19].

1.3 Genetics alone does not describe what is expressed – transcriptomics is necessary to quantify expression

The mutational burden in cancer cells represents the set of heritable changes occurring within cancer cells; a marker of the evolutionary process of the disease. However, this neglects to inform us of the phenotypic and functional implications of the evolutionary process on the cell, which is what serves as an ultimate determinant of the fitness of the cancer cell. As a consequence of natural selection, if a given cell is better adapted to survive in a given niche, this cell will generate a clonal lineage, which can be inferred from genomic sequencing data. However, in order to determine the selection processes by which these clonal lineages arise, it is imperative to understand what the functional effects of such mutations in cancer cells. Thus, while proteomics has not yet reached a large scale, over the past 10 years, technologies enabling the rapid and accurate measurement of the RNA transcriptome have become widely available. These, backed by many validation experiments, have led to a revolution in the understanding of the effects of mutations on the transcriptome. Increasingly, there is a scientific consensus

that tumours are not simply clonal populations of only a single cell type; rather, they are populations comprised of a heterogeneous group of cell types, with cells even potentially sharing genotypes, but differing among the genes that are expressed [20].

Insight at the level of the expressed mRNA within a given population of cells has enabled an understanding of tumour biology at a deeper level. Tumours with similar expression profiles of the cells have been identified by clustering to define tumour subgroups, as opposed to clustering based only on mutational information. Recent work has shown that these subgroups are highly clinically relevant, and often function well as prognosticators [21], risk stratifiers to determine the aggressiveness of therapy [22], and even diagnostic tests (e.g. cancerSeek) [23]. Through large-scale RNA-sequencing efforts such as TCGA project, these subgroupings have been identified, and have been used to define novel clinical sub-classifications representative of the vast heterogeneity within tumours, even from the same primary site. Improving the understanding of the differences between tumours and molecular subtyping has led oncology closer to the goals of personalised medicine [22, 24, 25].

1.4 The composition of the human transcriptome

The human transcriptome is defined as the sum of all ribonucleic acid components, transcribed from the underlying DNA contained within the cell [26]. The most abundant component of this is the ribosomal RNA component, which makes up approximately 80-90% of all RNA in human cells by mass [26]. These ribosomal RNA molecules are comprised of the various subunits that combine to create the ribosomal machinery by which protein products are translated from mRNA in the cell. Following ribsosomal RNA, the next most abundant species within the transcriptome is transfer RNA, representing 10-15% of all RNA by mass [26]. These molecules primarily function as shuttlers for individual amino acids to the sites of nascent protein production during active translation in the ribosome. The next most abundant RNA species is mRNA, which represents 3-7% of RNA by mass [26]. These three primary types of RNA comprise approximately 99% of all RNA within the cell, and loose estimates only exist for the remainder of the species. Of the remaining 1%, approximately 70% of this is thought to arise from pre-mRNA, small nuclear and nucleolar RNAs [26]. Of the remainder, miRNA are thought to comprise 0.003-0.02% of total cellular RNA by mass, circRNA are thought to comprise 0.002-0.03% of RNA by mass, and long non-coding RNA (lncRNA)s are thought to comprise 0.03-0.2% by mass [26]. That is, the RNA species that are at the focus of this thesis are in minority within the

transcriptome; forming a less than one-tenth of one percent of the mass of all cellular RNA at best.

1.5 Non-coding RNAs form a key component of the transcriptome, and have been shown to fulfill diverse roles

Through large-scale profiling of the transcriptome, many regions of DNA previously thought to be junk DNA, not coding for proteins, have been found to be expressed as RNA, termed ncRNA. These constitute a diverse family of evolutionarily conserved molecules, including lncRNA, circRNA, and miRNA, among others [27, 28, 29]. Much work has focused on the characterisation of non-coding RNA, including by the ENCODE consortium [30], has shown that these species, particularly miRNA, are involved in a number of developmental and differentiation-related processes [31].

1.5.1 miRNA repress mRNA through the RISC

Mature miRNA are small, single-stranded, linear, 22 nucleotide RNA molecules found within the cytoplasm of the cell that, in most cases, act to stop translation of mRNA or degrade mRNA [32]. Each miRNA acts on a specific set of mRNA targets through complementary base pairing between a seed region of the miRNA and a sequence at the 3' untranslated region (UTR) of the mRNA [33]. In a perfect match to this seed region, it is thought that the target mRNA is degraded, and in the case of an imperfect match, translation is slowed, resulting in less production of the protein product [34].

This repression is not achieved in isolation - miRNA must interact with a number of RNA binding proteins to form the RNA-induced silencing complex (RISC), mediating their repressive effects, as described in Figure 1.2A [34]. This enables repressive activity to occur, and is comprised of the miRNA itself and the protein Argonaute 2 (*AGO2*) primarily, as well as *AGO1*, *AGO3*, and *AGO4* [34]. The mRNA that are degraded are done so primarily by an RNA slicing domain in the *AGO2* protein, and the mRNA which are not degraded by this mechanism experience decay and delayed transcriptional activity whilst bound to the RISC [35]. mRNA decay occurs by *TNRC6A-C* proteins acting within the RISC complex, which recruit de-adenylating proteins to the RISC, such as polyA-binding protein (*PABPC1*) and the *CCR4-NOT*

deadenylase complex [36]. Once deadenylated, mRNA molecules may be degraded from the 5'-3' direction by *XRN1* exonuclease [34].

The mechanisms of translational repression for the target mRNA have been shown to be similar to the mechanisms of degradation, as the *TNRC6A-C* proteins in humans mediates the recruitment of translational repressors such as *DDX6* to the target mRNA [37]. Alternatively, there is a proposed interaction of the RISC and the elongation factors controlling translation initiation, resulting in the displacement of *EIF4A* from *EIF4F*, such that ribosomes are prevented from binding the mRNA being targetted, thereby inhibiting translation [38]. Additionally, deadenylation resulting from the removal of *PABPC1* from the polyA tail of the mRNA will also result in translational repression, as its loss will reduce the recruitment of elongation factors critical to encourage further ribosomal binding [34].

1.5.2 Specific miRNA may act to increase mRNA and protein levels

Interestingly, the opposite effect of miRNA mediated repression - namely, miRNA mediated activation, has also been observed in human cells. Vasudevan et al. in their work show that miR-369-3 binds to an AU-rich element in the *TNFα* gene [40]. This results in the recruitment of an *AGO2-FXR1* complex, which activates translation of the *TNFα*, thereby upregulating levels of this protein [41]. They suggest that miRNA may oscillate in their behaviour between repression and activation of mRNA translation, depending on timing within the cell cycle. Unexpectedly, they have shown that miRNA function, through complementary base-pairing and recruitment of ribonucleoprotein complexes to actively transcribed mRNA can both activate and repress translation, suggesting an important dual role that remains underexplored within the literature [40].

1.5.3 miRNA are produced via a canonical biogenesis pathway

The production of miRNA within the cell, a process referred to as miRNA biogenesis, is a well-ordered process of several steps, as outlined in Figure 1.2B. The production of miRNA begins in the nucleus, where RNA polymerase II transcribes the primary miRNA (pri-miRNA) transcript from the DNA [42]. These transcripts may be polycistronic, containing multiple miRNA if transcribed from miRNA clusters in the genome [43]. Next, this pri-miRNA is processed into a hairpin structure called a

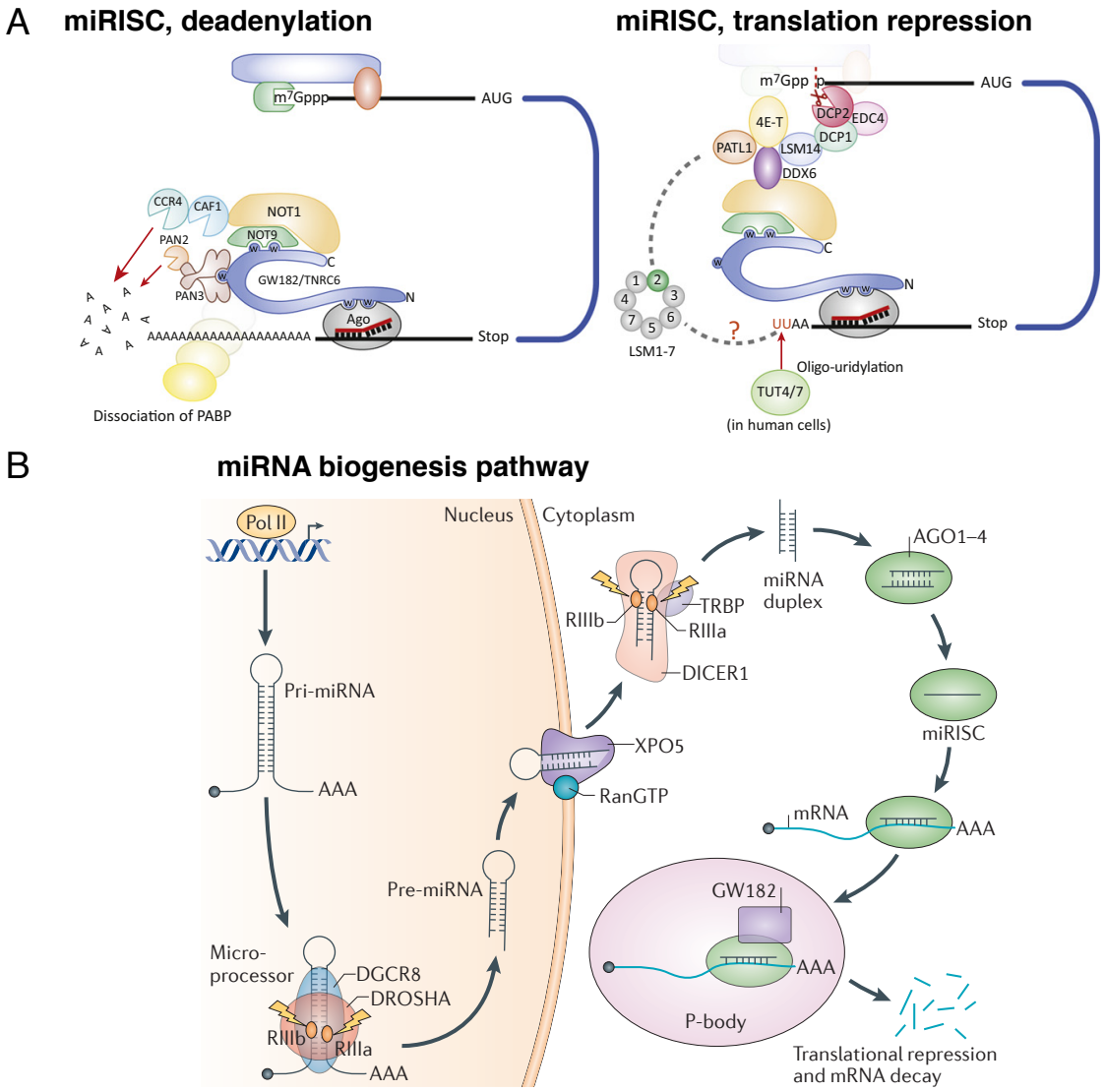

Figure 1.2: **miRNA biogenesis pathway, and the formation of the RISC (RNA-induced silencing complex).** Images adapted from Iwakawa and Tomari [34] (a) and Lin and Gregory [39] (b). a) A schematic of the proteins involved and structural relationships between them in miRNA-mediated repressive actions; deadenylation (left) and translational repression (right). b) The proteins involved in the canonical miRNA biogenesis pathway. Figures reproduced with permission.

precursor miRNA (pre-miRNA), by a complex of ribonucleoproteins termed the microprocessor complex [44]. The microprocessor complex consists of enzymes *DROSHA* and *DGCR8* in humans, which cleave the relatively long pri-miRNA transcript into a 70nt hairpin structure, the pre-miRNA [45]. The pre-miRNA hairpin structure is then exported out of the nucleus for further processing, through a nuclear pore comprised of the proteins Exportin-5 (*XPO5*) and Ran-GTPase (*RAN*) [46]. Once in

the cytoplasm, the pre-miRNA hairpin structure binds a complex consisting of the RNAse-III endonuclease *DICER*, *AGO2*, and *TNRC6* [47]. Dicer cleaves the hairpin loop pre-miRNA, resulting in a RNA duplex of two short, single-stranded sequences, which are the two arms of the mature miRNA (also termed the guide strand and the passenger strand) [48]. These are then loaded into the RISC, and are unwound so that one strand functions as the repressive pre-miRNA, as the guide strand [42]. It is then thought that *AGO2* functions to degrade the passenger strand of the miRNA duplex loaded within the RISC [49]. Once passenger strand degradation and RISC assembly with the guide miRNA is complete, it is able to bind to target mRNA, and achieve its repressive (or potentially activating) functions, as outlined above [33].

1.5.4 Technical aspects of next-generation RNA-sequencing

Transcriptomics in the current age has evolved into a science of high-throughput data generation, as a result of next-generation sequencing (NGS) technologies [50]. Here, I briefly describe the experimental approach for NGS using the Illumina protocol; the most widely used approach, and the one taken by the TCGA consortium, a dataset relied upon heavily in this thesis. NGS is conducted through three main steps: library preparation, sequencing, and data analysis [50]. The protocol for RNA-sequencing is described below, noting that whole genome and whole exome sequencing are similar, but with different library preparation steps.

Library preparation Library preparation refers to the part of the protocol encompassing extraction of the RNA from the samples of interest to the point at which complementary DNA, synthesised from this RNA with small oligonucleotide sequences called adapters ligated to the ends can be added to the sequencer. The first step in library preparation is the extraction of RNA from the population of interest, typically followed by a selection step to ensure that sequencing is only done for the RNA type of interest. Usually, this selection step is for polyadenylated RNA, through elution with a column containing beads lined with oligo-dT, enriching for primarily mRNA. Next, these RNA molecules are fragmented into smaller sizes, to be compatible for bridge amplification in the sequencer, a process described in further detail below. From these small RNA fragments, complementary DNA (cDNA) is synthesised, as DNA is more stable than RNA, within the sequencing platform. Subsequently, short DNA sequences, called adapter sequences, are ligated to the ends of these cDNA molecules. These sequences facilitate both sample identification (for mixed samples being sequenced in the same lane), and attachment to the flow-cell, the core element

of the sequencer. These DNA fragments are then amplified through a polymerase chain reaction (PCR), and are subsequently denatured into single-stranded segments, and are then ready for loading into the lanes of the flow-cell [51].

Sequencing Once the library has been prepared, as described above, it is ready to be loaded into the Illumina flow-cell. The flow-cell is a glass slide with multiple lanes, coated by a grid of vertically-oriented oligonucleotides, complementary to the adapter sequences ligated to the ends of the cDNA, as depicted in Figure 1.3 [52]. The sample library then hybridises to the vertical oligonucleotides on the flow-cell, in a sparse manner, ensured by a dilution step with buffer solution, before loading into the glass slide. Then, through a process known as 'bridge amplification,' clusters of oligonucleotides are sequenced in parallel. Bridge amplification consists of the oligonucleotide sequences physically bending over, and attaching to a nearby oligonucleotide on the slide through base complementarity, such that the two ends of the cDNA molecule are affixed to the flow-cell. Then, through a sequencing step, similar to PCR, a complementary strand is synthesised, and then the strands are denatured. In this way, 'clusters' of nearby oligonucleotides on the flow-cell all contain sequences from the same cDNA molecule in the library [53].

Following this, along the vertically oriented single-stranded molecules, nucelotide-by-nucleotide sequencing takes place, with each addition of a nucleotide resulting in the emission of light at a specific wavelength [50]. The signal is amplified as a result of the clusters of similar sequences arranged throughout the flow-cell, enabling the more accurate detection of the base added to each sequence [52]. A detector records this information as the raw 'reads' from the sequencer. An additional sequencing step starting from the opposite end of the cDNA molecules can also be done, and these reads can also be recorded, in a process called paired-end sequencing. This enhances the accuracy of the sequencing, and facilitates better resolution of repeated regions along the genome, as a more accurate picture of each molecule emerges [50].

Data analysis Once all of these raw reads have been generated, the next step is to align these reads to a reference genome. The reads consist of 50-100 base pair fragments, with a quality score for the confidence of the machine in identifying each nucleotide at each position of the read, and two reads if paired-end sequencing has been done [50]. These reads are then broken up into smaller fragments, each of which are aligned to the reference genome. Every segment of the reference genome can then be counted for number of reads spanning specific regions, through a number of

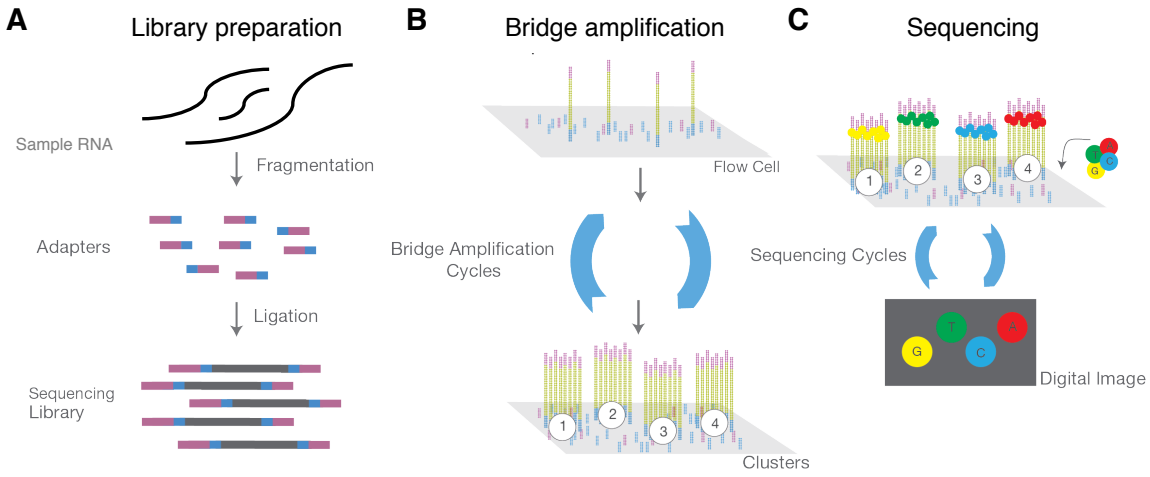

Figure 1.3: **Overview of steps involved in Illumina next generation sequencing.** Image adapted from Illumina promotional materials ('An Introduction to Next-Generation Sequencing Technology') [54]. a) Library preparation involves sample RNA reverse transcribed into cDNA, fragmented, and ligated with adapters. b) Sequencing library is amplified on the flow-cell using bridge amplification, generating clusters of oligonucleotides. c) Oligonucleotides are then sequenced one nucleotide at a time, creating a digital image, which is read to interpret the sequence of reads in a parallel fashion. This raw data is then processed by aligning reads to the genome, and determining counts. Images are reproduced with permission.

algorithms [55, 56, 57]. From these reads, the expression of each region comprising exons of the genome, splice variants, non-coding RNA, and how these vary across samples, can be inferred. Next, because of intrinsic differences between the samples, the number of reads must be normalised to make it comparable across samples. There are many methods of normalising expression from this raw data, but most common methods centre around the issue of different numbers of total reads for different samples in the sequencer, owing to loading differences, or differences in sequencing depth. That is, there may be different numbers of reads recorded, even if expression may have originally been similar, and the normalisation protocols seek to correct for this [58].

Limitations and issues with RNA sequencing While RNA sequencing has substantially changed the field of molecular biology, it is not without technical limitations. In particular, nucleotides may be read with poor quality, owing to the optics of the photodetector behind the flow-cell, as small signals may be obscured by the neighbouring amplified signals. More specifically, this occurs when cDNA libraries are not appropriately interspersed on the flow-cell. This leads to the signal from

a cDNA species being 'drowned out' by the clusters arising around it from bridge amplification [59]. Secondly, the identification of reads from long repeated regions of DNA remains an issue, as these are difficult to map from small and specifically single-end reads [60]. Moreover, different aligners and different versions of the human genome can lead to different counts for the same genes, which leads to issues of poor computational reproducibility in some cases, or different parameter settings for alignment software yielding different results for an experiment [58].

Small RNA sequencing enables identification of miRNA The initial preparation of the RNA from the sample of interest has the potential to heavily bias the sequencer to sequence molecules of interest. For instance, this is why polyadenylated transcripts are selected for when mRNA are of interest, as this will remove nearly all non-mRNA molecule from the library, ensuring that reads and regions of the flow-cell are not wasted on ribosomal RNA (rRNA) or other forms of RNA, which may be abundant in the sample [50]. In this thesis, I rely on results from miRNA sequencing, a form of small RNA sequencing. The process for miRNA-sequencing is largely analogous to that of RNA sequencing as described above, but with the added step of miRNA enrichment of the sample prior to the sequencing. That is, miRNA are selected for by a size criterion of the cDNA before or after adapter ligation, often by cutting out the band corresponding to miRNA-sized fragments from a gel electrophoresis [61]. In this way, the miRNA present within the transcriptome can be preferentially selected for sequencing based on a size threshold. Thus, if library preparation is to be done after size selection, library preparation may proceed via the steps of adapter ligation to the small RNA isolated, transcription to cDNA, and then the addition of multiplexing indices, followed by sequencing on a flow-cell.

1.5.5 miRNA nomenclature

There are many systems of naming miRNA that are in use throughout the scientific literature, however for the purposes of this thesis, I will ascribe to the following conventions, based on the current standard used at the time of writing by the miRbase database [62, 63]. Each mature miRNA in this study is from *Homo sapiens*, unless otherwise indicated, thus starting with the prefix hsa-, which I omit for brevity. Mature miRNA have started (typically) with the prefix hsa-miR-, indicating a mature miRNA from *Homo sapiens*. However, cases of historical naming, such as the hsa-let- family of miRNA are also present. Each mature miRNA comes either from the 5' or 3' strand of the precursor hairpin molecule, and this is indicated by the suffix

-5p or -3p, respectively. In this thesis, I do not use the older convention indicating lesser expressed miRNA transcripts with an asterisk. To indicate immature precursor miRNA strands, in this thesis I abide by the nomenclature with uncapitalised hsa-mir- as the prefix, and no suffix [63].

1.6 What can a computational study add?

The push into the genomic era has facilitated the generation of a vast amount of molecular data about tumours and cancer cell lines, across many tissue types, microenvironmental conditions, and clinical phenotypes. Within these rich datasets, the role of non-coding RNA remains relatively unexplored, despite experimental evidence for its important role in cancer. An understanding of the roles of non-coding RNA may ultimately lead to improved diagnostics, prognostics, and therapeutics for patients with cancer. In using these large datasets, there is the ability to see conserved patterns across vast numbers of tumour samples, allowing for the generation of novel hypotheses for small effects that are not easily observed from targeted experimental approaches. Moreover, in looking for these conserved patterns, I design novel approaches, uniquely integrating knowledge of genomics, clinical medicine, computation, and mathematics. That is, in this thesis, I synthesise, from these disparate fields, the methods that facilitate hypothesis generation for the roles of non-coding RNA in tumours, based on patterns only observable from large genomic datasets.

Having generated these hypotheses, I use theoretical modelling to explore the functional implications of these ideas. By designing and studying models and their behaviour *in silico*, I show how to generate and test hypotheses about model behaviours at a scale not possible by *in vitro* or *in vivo* experimental approaches. For instance, in this thesis, I show the range of possible dynamics occurring in a transcriptional feedback loop including a miRNA sponge, and the conditions where these dynamics occur. While theoretical models are not precisely representative of every facet of biological reality, when designed in a way that captures the most important phenomenological features, their results can be extremely useful, and informative for experimental validation. The models discussed in this thesis are those for which evidence is provided through associations demonstrated through large genomic datasets of tumour samples.

1.7 Scope of thesis and summary of presented work

In this thesis I aim to create an understanding of the roles of non-coding RNA, specifically miRNA and circRNA, as they interact with other molecules in the transcriptome in human cancers. I focus primarily on human cancers, and the phenotypes described by the hallmarks of cancer, with a special focus on the role of tumour hypoxia. The approach I take is data-driven, and initially informatics-based, leveraging the scale of large datasets to separate signal from noise, and elucidate functional mappings for these species of non-coding RNA. I take also an approach based firmly on mathematical modelling, specifically highlighted in the final chapter to better understand from a dynamical systems perspective, the role of miRNA sponges. This analysis has resulted in novel conclusions about how oscillatory behaviour can originate in the transcriptome of a cell, and how different types of miRNA sponges may be implicated for different functions. In this Introduction, I have discussed the preliminaries for understanding the work that this thesis rests upon in a broad sense, and note that within each chapter, I present a more focussed literature review and review of key concepts.

Thus, the following questions were investigated through this thesis:

- How can gene signatures be reliably assessed? (Chapter 2)
- How do miRNA associate with the hallmarks of cancer, and can this be assessed through an analysis of gene signatures? (Chapter 3)
- How is miRNA biogenesis associated with hypoxia? (Chapter 4)
- How are circRNA associated with cancer and tumour hypoxia biology? How might these associate with miRNA biogenesis? (Chapter 5)
- How might circRNA and other miRNA sponges affect dynamics of gene networks in the cell, and in which processes may these dynamics be useful? (Chapter 6)

Chapter 2

sigQC: A procedural approach for systematic evaluation of gene signatures

Abstract

With the increase in next generation sequencing generating large amounts of genomic data, gene expression signatures are becoming critically important tools to help interpret this data. As a result, gene signatures are poised to make a large impact on the diagnosis, management, and prognosis for a number of diseases. However, gene signatures have been limited in their wider application thus far because of a lack of quality control, resulting in poor reproducibility, especially in the clinical setting. In evaluating the quality of a gene signature, it is crucial to establish whether a signature can be used on a dataset, and if so, how the signature should be used. In this chapter, I discuss these issues in light of the statistical properties that a dataset and gene signature should possess to ensure reproducible behaviour. Having defined these properties, the first quality control protocol for gene signatures, called *sigQC*, is introduced. This has been designed to facilitate a streamlined, systematic approach for the evaluation of gene signatures across datasets. Ultimately, this protocol aims to increase the scope for using gene signatures by ensuring in a given dataset, the signal captured is reliable, and if not, how the signature may be refined.

2.1 Introduction

2.1.1 Gene signatures are increasingly being used in the clinic

Gene expression signatures, over the past decade have helped advance the understanding of disease, pathogenesis, and clinical response [64, 65, 66]. As the cost of sequencing decreases, the development of classifiers predictive of many types of outcomes may increase. However, because of the inherent technological differences in genomic sequencing technologies, and the differences in measurement and post-processing, the reproducibility of gene signatures and ensuring quality in their application remains an open problem.

Current gene expression signatures used in the clinic have shown utility in assisting with clinical decisions, often providing support by therapeutic stratification of patients. For instance, the OncotypeDX breast cancer gene signature uses a set of 21 genes to predict the risk of recurrent breast cancer in patients with estrogen receptor (ER)-positive, node-negative breast cancer, guiding the decision for adjuvant chemotherapy [67]. Similarly, the Mammaprint signature uses the expression of 70 genes to predict the risk of breast cancer recurrence in Stage I or II cases [68]. Both of these genomic tests are able to circumvent the technical differences in gene expression quantification by centralising all of their processes, and generating a single score classifier by an in-house algorithm. This in-house aspect to both of these tests limits the wider applicability, testing, and validation of these signatures, and raises the cost of such tests for patients. In order to facilitate the wider adoption of gene signatures in the clinical setting, it is necessary to determine which signatures can be used in a variety of settings, by reproducible scoring metrics.

2.1.2 Many types of gene signatures are possible

Before proceeding any further, the term gene signature for the purposes this chapter is defined as follows, in agreement with the Broad Institute, as a set of genes for which mRNA expression displays a consistent pattern in conjunction with a biological process, phenotype, or outcome [69]. That is, gene signatures as used in this chapter are differentiated from signatures such as mutational signatures [70] and multi-gene mutation panels [71]. This definition lends itself to the quality control approach presented within this chapter, termed *sigQC*. By analysing metrics on the patterns of expression of given gene sets on a dataset, the fairness of summarising such a set of genes into a single signature score is determined, or alternatively whether a more dataset-specific approach might be required.

Often gene signatures are sets of genes associated with an underlying statistical or linear model, facilitating the generation of a single numeric signature score from the full set of signature gene expression values for a given sample. The *sigQC* methodology developed is agnostic to this underlying model. Thus, the signature summary score is abstracted to relatively simple non-parametric measures of central tendency of a gene set; namely, the median, mean, and first principal component. As a result of this abstraction, *sigQC* is only appropriate for application to gene signatures in which all signature elements vary in the same direction (i.e. up or down) in association with the outcome or phenotype of interest.

2.1.3 Gene signatures may be derived in many ways

Gene signatures are derived by an ever-increasing arsenal of methodologies, spanning approaches such as supervised [72] and unsupervised clustering [73], seed-based approaches [74], and other machine-learning techniques [75]. Each of these approaches has its advantages and disadvantages in terms of the genes identified and the properties of the resultant signatures, but there are no standards for signature generation in different scenarios. In particular, machine-learning based approaches may lend themselves towards less transparent, more complex underlying statistical models, depending on the type of approach used [75]. Likewise, seed-based approaches, as used by Buffa et al. in [74] are those that inherently rely on co-correlation and co-expression of signature elements, and may be better suited to simpler summary statistics, such as the median expression of signature genes. Here, the approach used for the generation of a given gene signature is not considered, but rather the issue of generalisability after one has obtained a signature is addressed. That is, the input to this protocol is the signature itself, alongside a dataset on which the quality should be determined, thereby allowing for flexibility in how the signature is generated.

2.1.4 Gene signature-based analyses show relatively low reproducibility

In many cases, gene signatures remain limited to narrow use cases, or lack disease specificity in predictive power. In fact, it has been shown that random gene signatures are capable of statistically significantly separating groups of breast cancer patients with favourable and unfavourable outcomes [76]. In response to these results, Berglund et al. noted that the lack of specificity and reproducibility among gene signatures is due to poorly behaved gene sets. From this, Berglund et al. proposed a

framework by which signature coherence, uniqueness, robustness, and transferability are evaluated for principal components analysis (PCA)-based signatures, and showed how these may be applied to check for reproducible behaviour [77]. Further, reproducibility among independent validation cohorts remains a challenging limitation to the widespread adoption of many gene signatures. It has been proposed that gene signature validation in a clinical setting should involve prospective independent randomised studies, which are costly, time-consuming, and require intensive study design. The issues of reproducibility in the clinical use of gene signatures stems from technical differences between platforms and sites, and suffers from a reliance on signature scoring methods that are highly dataset specific. To address these issues, a methodology such as *sigQC*, functioning to evaluate gene signature quality before more general application of a gene signature to a dataset, may ensure that validation cohorts involving retrospectively generated data are fair and appropriate analyses.

2.1.5 General characteristics of reproducible gene signatures

The pertinent application of gene signatures to a vast array of clinical data depends critically upon the ability of the signature to perform robustly over a wide range of possible confounders. Noise, inter-platform differences for gene expression profiling, as well as sample collection method can all affect quantitative results and gene sets therefore require validation in independent datasets [78]. Moreover, a central problem that limits the applicability of gene signatures to narrow use cases is the difficulty in summarising the expression of a disparate set of genes into a robust and transferable single score for each sample. There are many methods to achieve this, but many of these methodologies are dataset or technology specific, and therefore limit the utility of a particular gene signature.

In order to ensure that the influence of such factors is reduced, several tests and validation criteria are proposed in this chapter. The aim of these is to empower the user to determine whether, for a given gene signature, the signature's statistical properties are conserved across datasets, and whether the signature may be summarised fairly into a single score. This technique is presented as a quality control protocol, to be used before applying a previously derived gene signature on a new dataset, and this is necessarily separate from a protocol for signature derivation. Each of the many methods of gene signature derivation has its advantages and disadvantages in different scenarios, and navigating these requires deep, domain-specific understanding. As such, the tests proposed define metrics to assist in determining whether an existing

signature can generalise to a new dataset. This approach also evaluates whether different datasets are similar with respect to a given signature.

Conceptually, this protocol, called *sigQC*, was designed to ensure that gene signatures are derived with characteristics suitable for clinical utility, and to elicit those properties that pertain to broader application. Whilst the protocol may be used to evaluate the statistical properties of any set of genes, it is particularly useful in those cases where the signature has been assembled so that the genes have co-ordinated expression with respect to a given phenotype. Examples of such signatures are meta-genes, frequently summarised as single-value score and used to rank clinical or biological samples based on a given phenotype, and signatures that have been generated for enrichment analyses (such as those available through molecular signatures database (MSigDB) [78]).

During the evaluation of a signature in a dataset, there are four key features that must be accounted for: i) signature technical transportability, ii) signature biological integrity, iii) signature suitability and iv) dataset suitability. These are reflected by the metrics in the *sigQC* protocol, and covered in more detail below.

Signature transportability Signature transportability refers to the use of a gene signature across datasets produced by different technologies, such as RNA-seq or microarrays, which quantify genes differently, though they may originate from the same sample. Over the previous decade, most gene signatures have been developed using microarray technology (i.e. a collection of complementary DNA probe sequences attached to a solid surface) but the majority of gene expression quantification at present is done by next-generation RNA sequencing, and signatures must be tested to ensure that they act the same way for both types of data. This is further complicated by the fact that microarrays themselves comprise a range of technical methodologies (e.g. spotting, in-situ synthesis) and may have different output characteristics (e.g. one-channel vs. two-channel detection) [79].

More specifically, because of the technical differences between RNA-seq and microarray, gene signature applicability can differ. For instance, genes expressed at relatively low levels are better quantified with RNA-seq, as their calculated expression using microarray-based platforms is skewed by the presence of background noise owing to non-specific hybridization [80, 81]. In addition, for genes with relatively high expressions, microarray-based methods may be unable to reveal the full extent of gene expression due to saturation effects, limited by the number of available probes

on the array [80, 81]. As such, one of the major differences between the two platforms is the ability of RNA-seq to more accurately quantify gene expression across a wider dynamic range, which may lead to unexpected performance of gene signatures derived through microarrays, or vice versa. The ability to detect gene expression across a greater dynamic range comes with the caveat of greater heteroscedasticity, as in RNA-seq datasets, genes with higher expression often show greater variance, necessitating more careful statistical analysis and preprocessing [80]. A further difference relates to the ability of RNA-seq to quantify structural or splice variants of transcripts, which themselves may be elements of a gene signature, as these may not be detectable in a microarray-based dataset if the complementary probe was not present *a priori* [80]. Through *sigQC*, by providing the user information about which elements of the input gene signature are present, and the expression and variance of the elements in the given dataset, these key aspects differing across platforms can be explored, and quality of application can begin to be ascertained.

Signature biological integrity Secondly, given datasets generated using the same technology, a signature's ability to represent a biological phenomenon in a general, reproducible behaviour in a specific context should be ensured, before moving on to wider application. To study the degree to which a signature is able to represent this heterogeneity, in *sigQC* the distribution of signature scores across datasets is quantified, and if applicable, this is done in conjunction with covariates describing the samples, if these are available. *sigQC* also includes an analysis of modality in the signature score distribution, and allows quantitative comparison of signature gene expression, visualised as heatmaps, to identify dataset or signature subsets showing distinct expression patterns.

Signature and dataset suitability Lastly, in the case of multiple signatures and multiple datasets for the same phenotype or biological process being captured, the signature under primary consideration should be the most suitable for both the dataset and the level of generalisability desired. When calculating a summary signature score, it is important to assess whether the metric summarising the signature score is an appropriate summary statistic for the given dataset, which is done in *sigQC* by comparing the co-correlation of scoring metrics. Moreover, in *sigQC*, for easy comparison across dataset and gene signature combinations, all quality control metrics, including these co-correlations, are summarised numerically and plotted on a radar plot. These

values are scaled to ensure comparability between the different cases considered, and this allows for the efficient comparison of many gene signatures across many datasets.

2.1.6 Structure of chapter

In this chapter the gene signature quality control metrics implemented by *sigQC* are motivated and introduced. In Sections 2.2.1 - 2.2.10, the steps of the protocol are summarised and it is highlighted how each of these metrics is able to identify well- and poorly-performing gene signature and dataset combinations. In Section 2.6, comparable methods for signature quality control are discussed, as well as the limitations of the *sigQC* protocol. Next, in Sections 2.4 - 2.5 sample cases of the quality control procedure are shown, highlighting how *sigQC* can help to identify issues with gene signature application, and how these may be addressed based on the outputs of the package. Specifically, Section 2.4 shows a case of comparing two gene signatures on a single dataset to evaluate which is a more appropriate fit, and highlights the differences that arise between these two signatures. In Section 2.5, a case of signature translatability is considered, to test how a gene signature can be applied on datasets quantified by two different platforms (microarray vs. RNA-seq), and whether these results are at all comparable.

2.2 *sigQC* protocol overview

2.2.1 General remarks

There are two overarching aspects to this protocol, the first being the tests of the properties of the genes comprising the signature itself, and the second being the properties of the dataset as it pertains to the signature genes. A flowchart of the procedure is depicted in Figure 2.1.

The evaluation of the genes composing the signature is primarily to determine whether signature genes cooperate to give a strong, coherent signal across the samples, which can be summarised into a single value. For clinical uses, signature genes must be above a reliable detection threshold, and must capture inter-patient heterogeneity, and therefore should be both expressed and varying. Furthermore, the distribution of expression values for signature genes within an individual sample should be coherent enough to summarise into a robust value for comparison across samples.

As important as the signature itself, are the statistical properties of the dataset to which it is applied. Thus, within this protocol, a search for structured subcomponents

of a gene signature or dataset is described, as in doing so, it can be determined whether there are subsets of genes or samples that could benefit from treatment as a distinct class. Finally, the *sigQC* package includes commands for bootstrapping, and evaluation of a set of negative controls, using both random resampling and gene label permutation, to reveal an understanding of the null distributions for each of the metrics considered in evaluating signature quality.

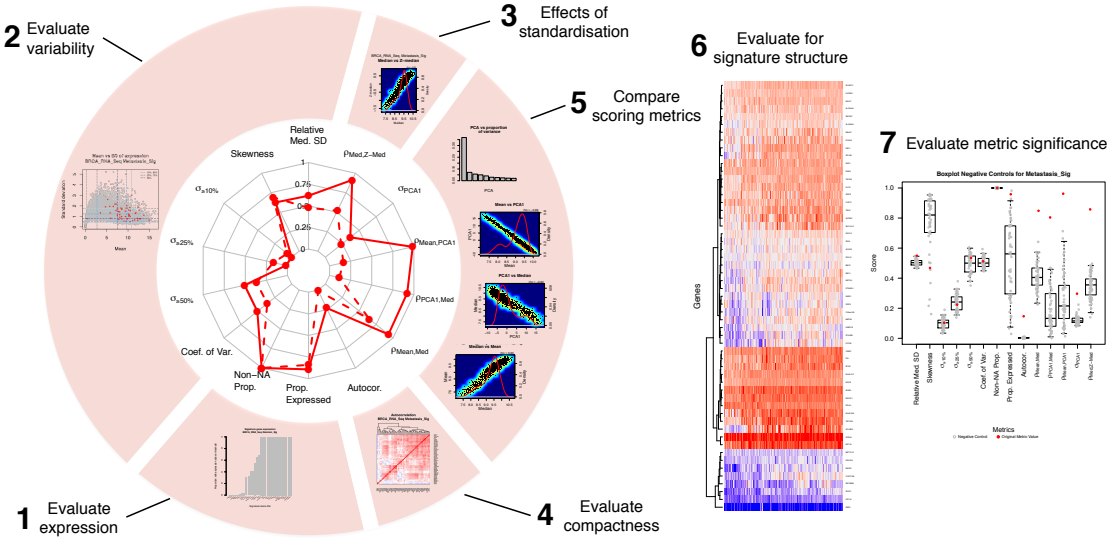

Figure 2.1: ***sigQC* protocol overview.** Flowchart of steps involved in the proposed *sigQC* protocol and sample output plots produced by the *sigQC* R package. Example outputs for a metastasis signature [82] on a clinical breast cancer dataset from the Cancer Genome Atlas Project [24, 83] are shown. The steps depicted are summarised as follows: evaluate expression of signature genes, evaluate variability of signature genes, interrogate the effects of standardisation, evaluate compactness (co-correlation) of signature genes' expression, compare metrics for scoring signature gene expression across samples, evaluate for signature structure, and evaluate each metric for significance by computing its null distribution.

2.2.2 (Un)-certainty in signature gene annotation

Prior to the testing of a gene signature, it is proposed to ensure compatibility between a gene signature and the dataset intended for use. In particular, because of a number of different annotation conventions for genes, the genes of a signature derived from one annotation of the genome should be able to be mapped to a matching annotation of the genome for the dataset under consideration, without loss of content or specificity. Several tools have been developed to accomplish this task; one widely used example is BioMart [84]. Because such mappings are generally not one-to-one, it is critical

to ensure that there is reasonable representation of all genes in a signature among the annotation used in a dataset of interest, as this uncertainty can detract from the functional ability of a given gene signature.

2.2.3 Evaluation of signature gene expression

A critical first step in the evaluation of the validity of a gene signature on a dataset is to ensure that the genes of the signature are expressed at a detectable level across the samples being considered, or at least in a sufficiently well large subset of samples. As a general rule, if genes within a signature are being used to differentiate biological or clinical groups, the expression value used to differentiate must be above a noise threshold. This threshold is context dependent. For example, in cases where lowly expressed genes may be key elements distinguishing biological or clinical states, a lower expression threshold is required, and this needs to be reflected by using assays that have less noisy measurements. In such cases, *sigQC* can aid by providing an indication on whether the minimal requirements of the assay are met. Additionally, a gene consistently not expressed or varying within a gene signature across a whole dataset contributes little to the overall use of the signature as a classifier, but this observation may be informative of the biology of the samples within the dataset. Thus, *sigQC* first evaluates the expression of all genes in the signature, and presents the proportion of samples expressing each gene at a supra-threshold level, as well as the proportion of all samples that have gene expression recorded as a non-NA value for each signature gene. The threshold for expression may be user-specified for each dataset, depending on the biological question asked, and on the technical characteristics of the dataset, such as the platform used. To aid this inspection, a graphical representation of this in the form of a bar chart and density plot showing the proportion of samples expressing each gene above a particular threshold is returned.

2.2.3.1 Data preprocessing

The importance of data preprocessing prior to applying a gene signature on a dataset cannot be understated. In evaluating the expression of genes of a given signature, it is crucial to ensure first that technical differences between the sequencing platforms have been corrected as well as possible, through normalisation. Depending on the technology available, there are multiple modes of normalisation, such as fragments per kilobase million (FPKM) or transcripts per million (TPM) counts in RNA-seq, or techniques such as robust multiarray average (RMA) for microarrays. Moreover,

if multiple datasets are used, it is recommended that these be treated separately for the purposes of gene signature characterisation, as there may be inherent statistical differences in the gene expressions computed between them. However, if combining these datasets into a single entity is crucial, batch correction should be applied; in essence, applying a statistical model to account for measured difference between the samples among the datasets.

Once normalisation has been assured, it is also imperative to ensure that the expression of all genes reported is on a scale that facilitates comparison using the gene signature. For instance, metrics of gene signatures that rely on measures of central tendency may be skewed by extremal values. As such, values for gene expression on a less extreme scale should be considered, such as log-transformed counts, when relying on gene signature metrics of central tendency.

Finally, depending on the type of statistical testing desired between gene signature scores, a transform to control for heteroscedasticity (a difference in underlying distribution of gene expression for each gene), may be necessary. In particular, for data from RNA-seq experiments, gene expression is known to show heteroscedasticity, in that genes with greater expression show greater variance, and log transforms or variance-stabilising transforms (VST) through R packages such as limma may better control for this effect [85].

While there is no one pipeline or set of transformations that can be universally recommended for data preprocessing, the principals discussed above are crucial considerations. This information is carried through the expression-variance plot produced through *sigQC*, wherein the effects of the data preprocessing as chosen by the user may be analysed for comparability, obvious batch effects, and heteroscedasticity.

2.2.4 Evaluation of signature gene variability

In addition to having non-zero expression across a number of samples, signatures that aim to stratify or classify samples should vary above the noise threshold across samples. Noise may occur in gene quantification for a number of reasons, including technical variability, non-specific hybridisation in microarrays, or issues with sample preparation. Thus, when any set of genes' expression is considered, it is crucial to ask whether the observed expression is due to noise, or whether some large component of it can be attributed to noise in the dataset. *sigQC* addresses this by ensuring that the expression of the gene in question is varying at a sufficient dynamic range of expression to ensure that its expression is not attributed to noise. More specifically, an evaluation step is done which involves the comparison of a standardised metric of

variance, the coefficient of variation, among the genes of the signature to all genes recorded in the dataset. To facilitate inspection, this result is provided both as numeric tables and as a scatter plot visualization of mean versus standard deviation for all genes, and their associated quantiles for mean and standard deviation, overlaid with the same scatter plot for all signature genes.

In the case of a signature performing well on a given dataset, the genes of the signature will be highly expressed and highly variable, as evidenced by the plots of expression and variability in Figures 2.2a-b and 2.3a-f. As shown in Figure 2.2a-b, the red dots, corresponding to the genes of the signatures are enriched higher-expression and higher-variability regions of the plot for the metastasis signature, as compared to the random gene set.

2.2.5 Effects of data standardisation

A subsequent issue with the application of gene signatures is the effect of data standardisation. A given signature may be applied on a set of data standardised in a one way for biomarker discovery, but this data is often re-standardised in another way for signature application. To account for this, it is proposed that the gene signature metrics and summarisation provided by *sigQC*, and illustrated in the following sections, should be compared using both non-standardised data and standardised data. In this way, the effect of gene expression standardisation on signatures which had been originally developed using absolute expression rather than standardised expression can be established. Likewise, it can be determined whether the information carried in the standardised expression is at risk of being lost when using non-standardised data.

2.2.6 Evaluation of signature compactness

A compact gene signature is one that contains genes with high levels of pairwise correlation, termed intra-signature correlation or autocorrelation. Often, this is the implicit assumption underlying the application of a signature. For example, gene set enrichment analysis [86] and related methodologies, widely used both in basic and clinical research, generate biological hypotheses on a given dataset based on the coordinated behaviour of genes within a set representative of a specific biological phenotype. In such cases, the gene signature with components acting in a co-ordinated manner ensures that the signature has more likely captured the biological phenotype of interest, and that summary scoring metrics will not have outliers detracting from

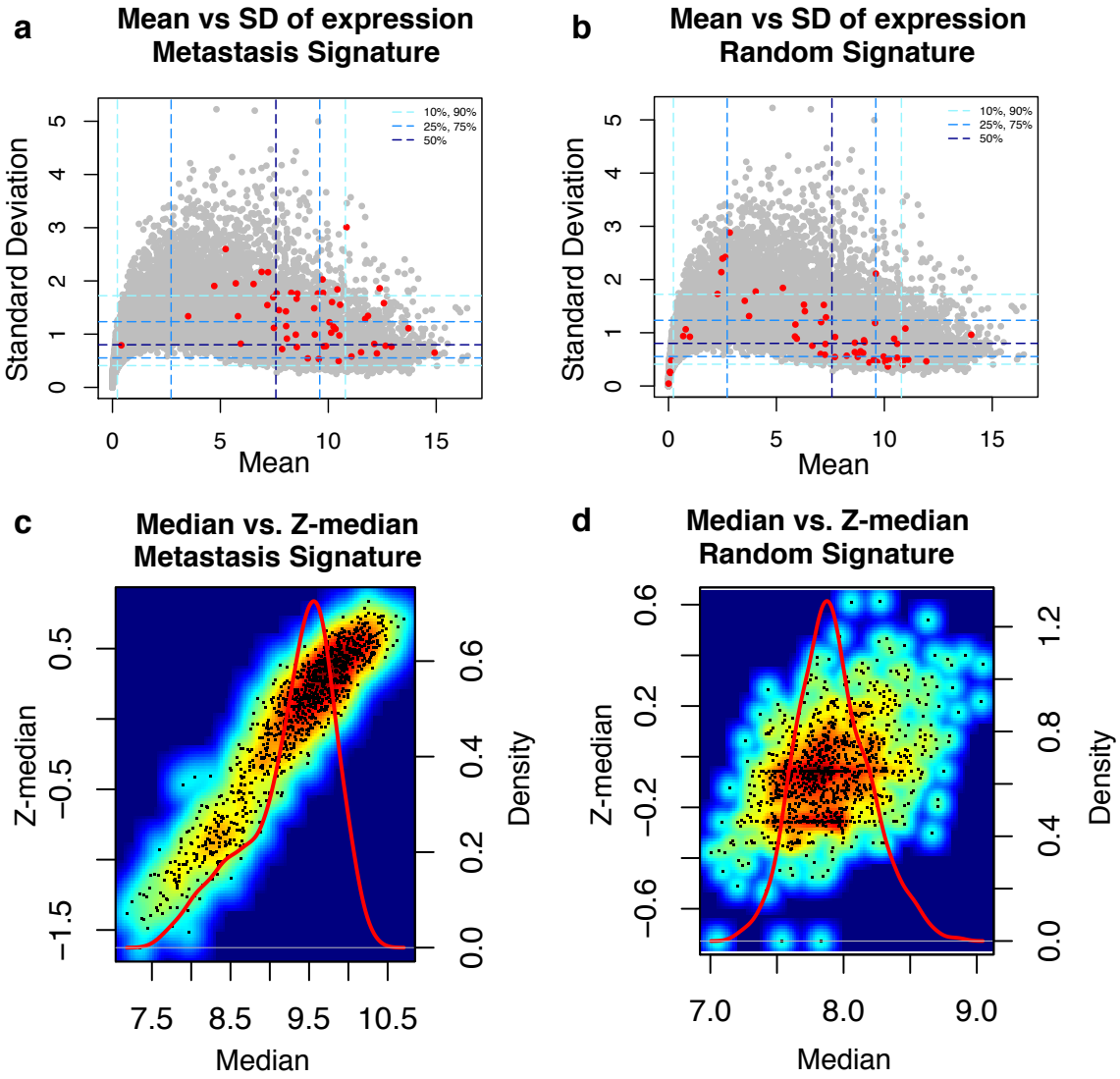

Figure 2.2: **Metastasis signature genes tend to be highly expressed and highly variable, and scores are not affected by standardisation.** Expression of signature genes and variability across datasets for RNA-seq breast cancer for the metastasis signature (a) and a random gene signature (b). Comparison of median and z-transformed median of signature gene expression across the RNA-seq breast cancer dataset for the metastasis gene signature (c) and the random set of genes (d).

the other genes of the signature. While this is often done as a step in the derivation of the gene signature, it is also proposed that this should be verified as a property that holds true in testing a gene signature prior to application. In doing so, this ensures that a key metagene property holds in the new dataset. Additionally, this can inform whether a given metagene is suitable for an application different to that originally envisaged, which is often the case with gene set enrichment approaches, and can provide an indication for further refinement or *de-novo* derivation. For example, a metagene

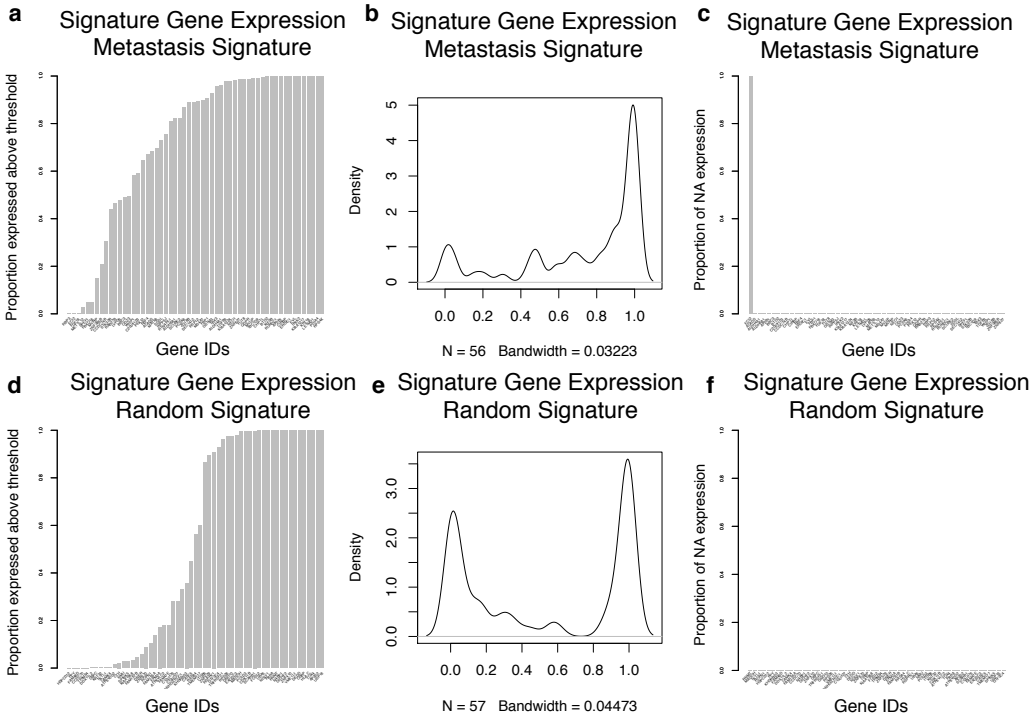

Figure 2.3: **Comparison of expression of signature genes reveals differences between metastasis signature and random signature in BRCA dataset.** Expression of signature genes across the TCGA breast cancer RNA-seq dataset for the metastasis gene signature (a-c) and a random set of genes (d-f), shown as (a, d) a barplot for the proportion of samples expressing a gene above the median, (b, e) a density plot showing the same information as the barplots in (a, d), and (c, f) a plot of the proportion of samples showing NA expression for each of the genes of the signature.

derived for a disease subtype might not behave as a compact set of genes when applied to another subtype of the same disease. However, understanding whether, or to what extent, the metagene behaviour is conserved can assist in the design of further studies and inform the biology itself. This verification step reveals signature genes or subsets of genes displaying a discordant behaviour in a new dataset, and also those genes which do maintain a compact behaviour across a number of normal and disease conditions. To test the level of intra-signature correlation among signature genes, in the *sigQC* package, the intra-signature correlation matrix is provided for the user, and a heatmap of correlation coefficients is created which compares the expression of every gene with every other gene in the signature.

2.2.7 Summarisation of signature gene expression to a single score

Across many domains of application for gene signatures, it is often desired to determine whether a signature can be summarised into a single score to enable the comparison between biological or clinical samples. The purpose of such a score is to encapsulate information from the entire signature, but to not be biased by outliers in the signature genes, which may detract from its performance. Such summarisation must be applied carefully as it may result in artifacts and erroneous conclusions if the signature genes did not have the requisite statistical properties for the score to be robust. To assess the suitability of such summarisation for a given signature, *sigQC* compares different score metrics; namely, a mean score, a median score and a principal component analysis-based, PCA score. Other metrics, or combinations of these metrics could be used, however initially these three metrics are suggested as they provide crucial basic information that can be used for designing more nuanced metrics.

The mean score, namely a score based on the arithmetic mean of the expression of the signature genes in each sample, is attractive for its simplicity. However, by using this score it is implicitly assumed that the mean is a fair summary for the distribution of the expression of all signature genes. This is not the case, for example, if the signature is not compact, as big expression changes could occur in opposite directions for different groups of genes without affecting the mean expression. Another case where a mean score would not be appropriate is in the case of skewed distributions. In such cases, outliers, namely a small number of genes expressed at higher or lower levels than the other genes in one or more samples, could heavily shift the mean score, and the subsequent ranking of the samples.

The median score, namely the median of the expression of the signature genes in each sample, is attractive for its robustness to outliers. This is a non-parametric score providing us with an indication of how the midpoint of the distribution of signature gene expression (the point dividing the population of genes in half based on their expression) changes across samples. The advantage of such a score is that the median expression is usually fairly stable, and it moves only if the expression of a substantial proportion of the genes changes coherently. However, similarly to the mean score, it may not be appropriate where the signature is not compact, as big expression changes could occur in opposite directions for different genes without changing the median, and it can miss subtle changes in expression of subsets of genes which might be biologically or clinically important.

Finally, PCA projects a set of observations of possibly correlated expression values into a new gene expression space of linearly uncorrelated variables, the principal components (PC). The transformation is defined such that the first PC accounts for as much of the expression variability as possible, or, in other words, has the largest possible variance. Then, each of the following PC has in turn the highest possible variance, under the constraint that it is uncorrelated to the preceding PC. By considering, in each sample, the magnitude of the first PC projection as a score, it is guaranteed that the score represents the change in the direction of the largest variation of expression. However, it is also implicitly assumed that this variation is sufficient to be a meaningful representation of the biology of the gene signature, which is a property of the input dataset. When this is not the case, the first PC fails to represent a large proportion of the variance, either further PC should be considered, or more complex summarisations approaches should be used.

Three further signature scoring metrics have also been built into *sigQC*, which were previously developed for signatures such as metagenes or those taken from repositories such as MSigDB [78]. These metrics are the gene set variation analysis (GSVA) algorithm [87], the single sample gene set enrichment analysis (ssGSEA) algorithm [88], and the pathway level analysis of gene expression (PLAGE) algorithm [89]. GSVA functions on the basis of calculating an enrichment score for a given gene signature (or gene set), and then computing the relative activity of these genes across samples, summarised into a score value for each sample [87]. ssGSEA calculates a gene set enrichment score, based on the gene set enrichment analysis (GSEA) approach for each sample, and this enrichment score is taken as the sample score [88]. PLAGE relies on first standardising the data using z-scores, and then uses the singular value decomposition in order to determine the weighting of each sample's expression on the given gene signature, as its score [89].

sigQC computes summarisation scores, providing their values and distribution, and asks whether the order of the samples is conserved when different scores are used to rank them. Score values are compared using Spearman correlation, that is the correlation of the samples' ranks. A high correlation between a mean and median score indicates that outliers, if present, have a contained effect. A high correlation between the median and PCA scores, indicates that the expression of the signature genes is changing in a coherent fashion (the signature is compact), and that these changes are well-represented by changes in the midpoint of the distribution. A high degree of correlation between all metrics gives a first indication that the signature has favourable properties to a single-score summarisation, and the relative ranking of

the samples is robust when different summary metrics are chosen amongst the ones presented.

2.2.8 Searching for structure in signature gene expression

Signature structure can be thought of as an underlying set of components comprising the signature, that tend to cluster together in terms of either gene co-expression, or groups of similar phenotype. Like the evaluation of signature compactness, structure of the signature and datasets are both taken into account during the development of a gene signature, but should also be verified in new datasets to ensure a similar pattern of gene signature expression. Furthermore, a change in the structure of a gene expression signature in a new dataset or context can reveal important insight into different technological issues or biological aspects, generating new hypotheses that can be tested.

Structure can be evaluated using various techniques; here the PCA is proposed (as introduced in the section above), as well as hierarchical clustering, for their easier visual interpretability with respect to other methods. This initial qualitative assessment is useful to prompt the need for further more advanced analyses of the signature structure. For example, this can be used to assess the level of redundancy present in a signature; that is whether different subgroups of genes carry similar information. Depending on the application, redundancy might be a sought-after, or an unwanted characteristic. Conversely, understanding whether independent subcomponents of a signature exist is an important part of evaluating a gene signature, as such components may signal biologically distinct sets of genes or samples within the datasets considered.

2.2.9 Comparison of multiple signatures and datasets

The *sigQC* package has been designed with an extensible framework, and can be used for the evaluation of multiple signatures and datasets at once. A summary plot produced by the package displays a host of metrics summarising the previous steps on a single radar plot. The elements described on the axes of this radar plot are summarised in Table 2.1, below. This visualization facilitates comparison of various metrics of multiple signatures on multiple datasets at once, with a single graphic image. Using this, the quality of various signatures, and the reasons for differences in quality, can be rapidly assessed over multiple datasets. Files including raw data of all

the statistics and summary scores computed by *sigQC* are also provided as output, for further analyses.

Metric Abbreviation	Metric Description	Metric Calculation
Relative Med. SD	Relative median standard deviation of signature genes as compared median standard deviation of all genes.	Consider the standard deviation of all signature elements' expression across all samples, then consider the median of this list, α. Similarly consider the median of the standard deviation of all reported genes across all samples, β. Value considered is $\lvert\alpha/(\alpha+\beta)\rvert$, where $\lvert.\rvert$ represents the absolute value.
$\rho_{Med.,Z-Med.}$	Absolute correlation coefficient of median of signature genes and median of signature genes on z-transformed dataset.	Absolute value of Spearman correlation coefficient between median and z-median of signature elements, used as scoring metrics across samples.
$\rho_{Mean,PCA1}$	Absolute correlation coefficient of mean and first principal component of signature genes.	Absolute value of Spearman correlation coefficient between mean and first principal component of signature elements, used as scoring metrics across samples.
$\rho_{PCA1,Med.}$	Absolute correlation coefficient of first principal component and median of signature genes.	Absolute value of Spearman correlation coefficient between first principal component and z-median of signature elements, used as scoring metrics across samples.
$\rho_{Mean,Med.}$	Absolute correlation coefficient of mean and median of signature genes.	Absolute value of Spearman correlation coefficient between mean and median of signature elements, used as scoring metrics across samples.
Autocor.	Median of autocorrelation values for all signature genes.	Median of list of all correlation coefficients for each signature element with every other signature element.
Prop. Expressed	Median proportion of samples expressing signature genes above threshold.	Median value of list of proportions of samples expressing each signature element above threshold for each signature element. Threshold is defined as median of expression of all genes, if not user-specified.
Non-NA Prop.	Median over all samples expressing each element as non-NA.	Median value of list of proportions of samples which have expression not recorded as NA, for each signature element.
Coef. of Var.	Median coefficient of variation of all signature genes, relative to the median coefficient of variation of all genes.	Consider the coefficient of variation of all signature elements across all samples, then consider the median of this list, α. Similarly consider the median of the coefficient of variation of all reported genes across all samples, β. Value considered is $\lvert\alpha/(\alpha+\beta)\rvert$, where $\lvert.\rvert$ represents the absolute value.
$\sigma_{\geq 50\%}$	Proportion of signature genes in the top 50% of all varying genes.	This is the proportion of signature elements that have coefficients of variation in the top 50% of all coefficients of variation for all genes.
$\sigma_{\geq 25\%}$	Proportion of signature genes in the top 25% of all varying genes.	This is the proportion of signature elements that have coefficients of variation in the top 25% of all coefficients of variation for all genes.
$\sigma_{\geq 10\%}$	Proportion of signature genes in the top 10% of all varying genes.	This is the proportion of signature elements that have coefficients of variation in the top 10% of all coefficients of variation for all genes.
Skewness	Relative skew of distribution of signature gene expression over all samples compared with skewness of overall expression distribution for all genes.	Consider the skewness of the distribution for the mean expression of all signature elements across all samples, α. Similarly consider the skewness of the distribution for the mean expression of all genes across all samples, β. Value considered is $\lvert\alpha\rvert/(\lvert\alpha\rvert+\lvert\beta\rvert)$, where $\lvert.\rvert$ represents the absolute value.
σ_{PCA1}	Proportion of gene signature score taken by median, by first principal component.	This is the proportion of the variance of gene signature score that is explained by the first principal component of the expression of the signature genes taken across all samples.

Table 2.1: **Description of metrics defining components of summary radar plot.**

2.2.10 Evaluation of null distribution of gene signature quality control metrics

Each of the metrics presented on the summary radar plot is computed for a given gene signature on a particular dataset, but to gain a greater understanding of the significance of these values, it is critical to consider the underlying null distribution from which each of these metrics arise. That is, for each dataset and gene signature combination, random resampling is performed to evaluate the underlying null distribution (or negative control) for the statistics considered. Namely, the distribution which would be observed under the assumption that there was no effect (e.g. the genes in a given signature were not correlated, or scores were not correlated). This is done using two different, widely-used approaches.

The first approach, random resampling, considers random gene sets with the same number of genes as in the signature evaluated. For each dataset and for the randomly resampled gene sets, each of the fourteen metrics considered on the summary radar plot is recomputed. These values comprise the null distribution for each of the metrics, and are useful in evaluating the significance of each of the quality control metrics. When using this technique, because gene sets are selected uniformly at random from all genes in a dataset, for gene signatures defined from more restricted subsets of genes, it is possible that the significance may have been artificially inflated by including irrelevant genes in the process of resampling. Thus, for a more nuanced calculation of the p value for significance of each metric, it is important to also consider the null distribution derived only from set of genes which were under consideration when the signature was originally derived, if this information is available.

The second approach is based on permutation resampling. Namely, instead of resampling from a random set of genes, resampling is done by randomly exchanging labels of the signature genes for each sample in each dataset. This provides a potentially stronger estimation of the null distribution for some metrics, such as the intra-signature correlation and the PCA, and has been previously used for similar analyses, such as gene set enrichment [90].

In addition, as a further evaluation of signature metric significance, it is proposed that a positive control gene signature, with known behaviour should be used as a comparator, when available. If there is a gene signature that has already been derived as a metagene on the given dataset, then this can be simply added to the list of gene signatures to be tested on a given dataset, and compared to the signature of interest to understand further how the metrics differ. While this positive control signature

may not always be available, when it is, it can be used as an important tool to better understand signature performance on the *sigQC* metrics.

2.3 Materials

2.3.1 Data sources

The remainder of this chapter has been written to highlight the utility in showing different use cases of *sigQC* through illustrative examples. In the first example case, two gene signatures are considered. The first is a relatively well-performing gene signature derived by Van't Veer et al., and reported in MSigDB [64]. This gene signature is comprised of 57 genes, and was derived using supervised clustering on a dataset of 117 breast tumours, with gene expression quantification by microarray, and the outcome of interest was distant metastasis at 5 years [64]. The signature derivation relied upon a three-step approach, wherein the authors first defined two groups of samples, one with distant metastases less than 5 years after presentation, and another without [64]. From these two groups, the authors identified 5000 statistically significantly differentially expressed genes, and from this, identified 231 that showed moderate-strong positive or negative correlation coefficient with the outcome [64]. From these 231 genes, the number of genes in the final prognostic classifier was optimised by reconsidering them in sets of 5, and using leave-one-out cross validation with the outcome of interest, identifying the optimal model, which resulted in the selection of 70 probe sets, representative of 57 distinct genes [64].

The second gene signature can be considered a dummy gene signature, and is 57 randomly selected genes. This signature was defined in this way to have a random comparator gene signature against which to compare the Van't Veer breast cancer metastasis signature.

The primary data sources considered for the work presented in this chapter were the breast cancer cohort from the TCGA, and a microarray dataset obtained through the gene expression omnibus (GEO).

The TCGA is among the largest compendiums of cancer data, collected across 11,000 patients in the US and Canada from over 20 institutions, comprising 33 different tumour types, with many samples having associated clinical metadata [91]. The sequencing of genomic material from these tumour samples and adjacent normal tissues taken from surgical resection was done in centralised locations to ensure good reproducibility of the results [91]. This data was generated through sequencing of 735 breast cancer samples, collected from newly diagnosed patients with no

prior treatments during surgical resection [83]. Frozen samples were examined to ensure that they contained less than 20% necrosis, per TCGA study requirements [83]. RNA was extracted from these samples using a Qiagen AllPrep kit, and the RNA preparation was sequenced using the Illumina HiSeq v2 platform, and processed data was accessed through the Firebrowse portal [83]. More specifically, RNA-seq by expectation-maximisation (RSEM) normalised mRNA gene expression for invasive ductal carcinoma of the breast were downloaded from the Firebrowse database at http://www.firebrowse.org, accessed on November 11, 2017.

In the second example considered, a microarray-based dataset was considered, obtained through GEO with accession GSE3494, accessed on December 1, 2017. This data series, published in 2005, was created from the analysis of frozen tissues of 251 patients with invasive breast cancer, collected in Sweden from 1987-1989, from which RNA was isolated from samples using the RNEasy Mini kit, with RNA evaluated for expression using the Affymetrix U133A and B arrays [92]. These data were pre-processed and normalised using the global mean method, and probe-set values were log-transformed and scaled to have mean target signal as log 500, per the authors' protocol [92]. The authors also repeated hybridisation on new arrays for samples with poor average signal intensities, as evidenced by elevated scaling factors (greater than 3.5), or poor signal for *GAPDH*, and also repeated hybridisation for arrays where visual artifacts were present [92].

2.4 Example use case: a comparison of two signatures

In this section, a sample use case of the methodology for gene signature quality control as implemented through the *sigQC* package is presented. Here, two gene signatures are considered on the same dataset, and it is shown whether one is a better fit for the dataset than the other. To highlight the differences between a well-performing signature and a poor performing signature, as described above, the Van't Veer breast cancer metastasis signature taken from MSigDB, consisting of 57 genes, was compared to a gene signature consisting of 57 randomly selected genes [69, 82]. The dataset on which each of these signatures was applied is the TCGA RNA-seq, invasive breast carcinoma dataset, accessed through the Firebrowse portal, as described above.

Analysis of expression: First, expression of signature genes was evaluated across samples in both datasets, and this was done by analysis of the plots shown in

Figure 2.3a-f. These plots describe the proportion of samples with supra-threshold expression of each signature gene, and the proportion of samples with non-NA values, identifying non-expressed signature components. The threshold used to define expression in this context was the median expression of all genes across all samples. From Figure 2.3a-f, it is clear that a greater proportion of the genes of the metastasis signature are expressed above the pre-defined threshold, as compared to the random signature. Additionally, these plots show that one of the metastasis signature genes is consistently expressed as an NA value in nearly all samples, and could be removed in future iterations of the signature.

Analysis of variability: An analysis of variability as produced by the *sigQC* package is shown in Figure 2.2a-b. These plots describe the mean and standard deviation of expression of all genes reported (in grey) versus all signature genes (in red), with corresponding dashed lines over the plots describing the 10th, 25th, 50th, 75th and 90th percentiles of both mean and standard deviation. This facilitates the identification of those signature genes which are not variable or expressed among the samples, as well as a global evaluation of signature behaviour across samples of a dataset. In this case, the genes of the metastasis signature (red dots, left plot), display higher expression and variability than those of the random gene signature, as demonstrated by their positioning closer to the right upper corner of the plot.

Analysis of data standardisation effects: An analysis of data standardisation effects is presented in Figure 2.2c-d. This plot provides the comparison of median of gene signature expression on the raw data provided versus the median of the gene signature expression on the z-transformed (standardised to zero mean and unit variance) dataset, for each sample in each dataset and each gene signature under consideration. The correlation between the unstandardised and standardised signature score is observed to be much greater for the metastasis signature, as compared to the random signature. This reveals that the standardisation of the score for the metastasis signature has little effect on the ranking of samples by signature score, which suggested that standardised expression values could be used with this signature, but not with the random signature.

Analysis of signature compactness: Next, heatmap and density plots of the correlation of each signature genes' expression with the expression of every other signature gene (pairwise correlation) were analysed, which revealed each signature's

compactness. These plots are presented for these signatures in Figure 2.4a-c where it can be seen that as expected, the breast cancer metastasis signature shows a high degree of compactness, and the random gene signature does not.

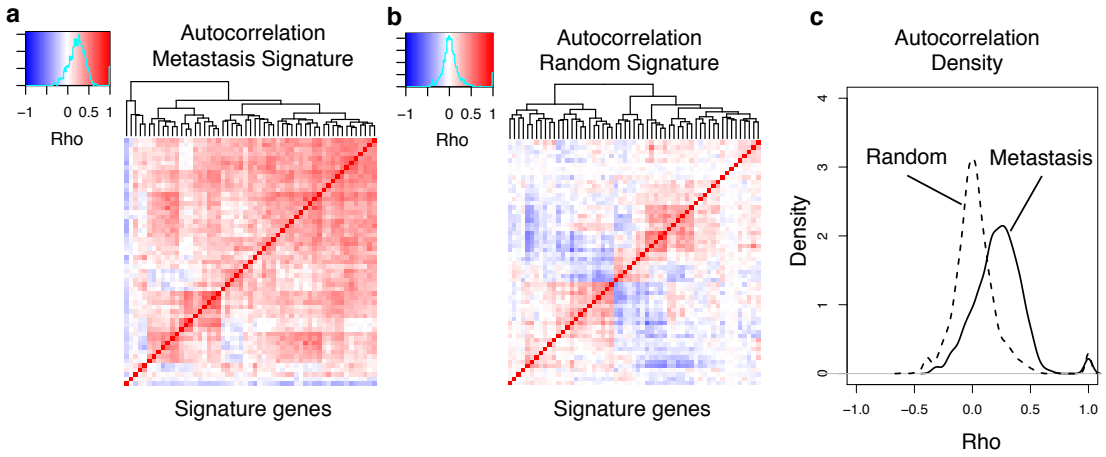

Figure 2.4: **Intra-signature correlation heatmaps revealed that the metastasis signature is compact.** Intra-signature correlation across datasets for a metastasis gene signature (a) and a random gene signature (b), with heatmap values represented in density plot form (c) for the TCGA breast cancer RNA-seq dataset.

Analysis of co-correlation of scoring metrics: Next, the pairwise correlation of mean, median and PCA1 as scoring metrics was analysed across the samples for each signature across each dataset, as depicted in Figure 2.5. Additionally, shown in the fourth row of panels of these plots in Figure 2.5 is a PCA scree plot, which describes the proportion of the variance attributable to each principal component, reflecting whether the first principal component represented a reasonable summary metric for a particular gene signature. For the metastasis signature, these scoring metrics all correlate well using the Spearman correlation ($|\rho| > 0.9$ in all cases), and the first principal component carries a large degree of the variance for the gene expressions. In contrast, for the random gene signatures, these scores do not correlate as well ($0.47 \leq \rho \leq 0.7$), and so care must be taken when deciding an appropriate scoring metric. Moreover, for the random signature, the first principal component represents a smaller proportion of the variability in gene expression, suggesting that it is not necessarily a strong choice of scoring metric.

Analysis of signature structure: Signature structure was evaluated considering the output of the *sigQC* package. First, signature structure was evaluated by hierarchical clustering on the provided expression values of the signature elements over

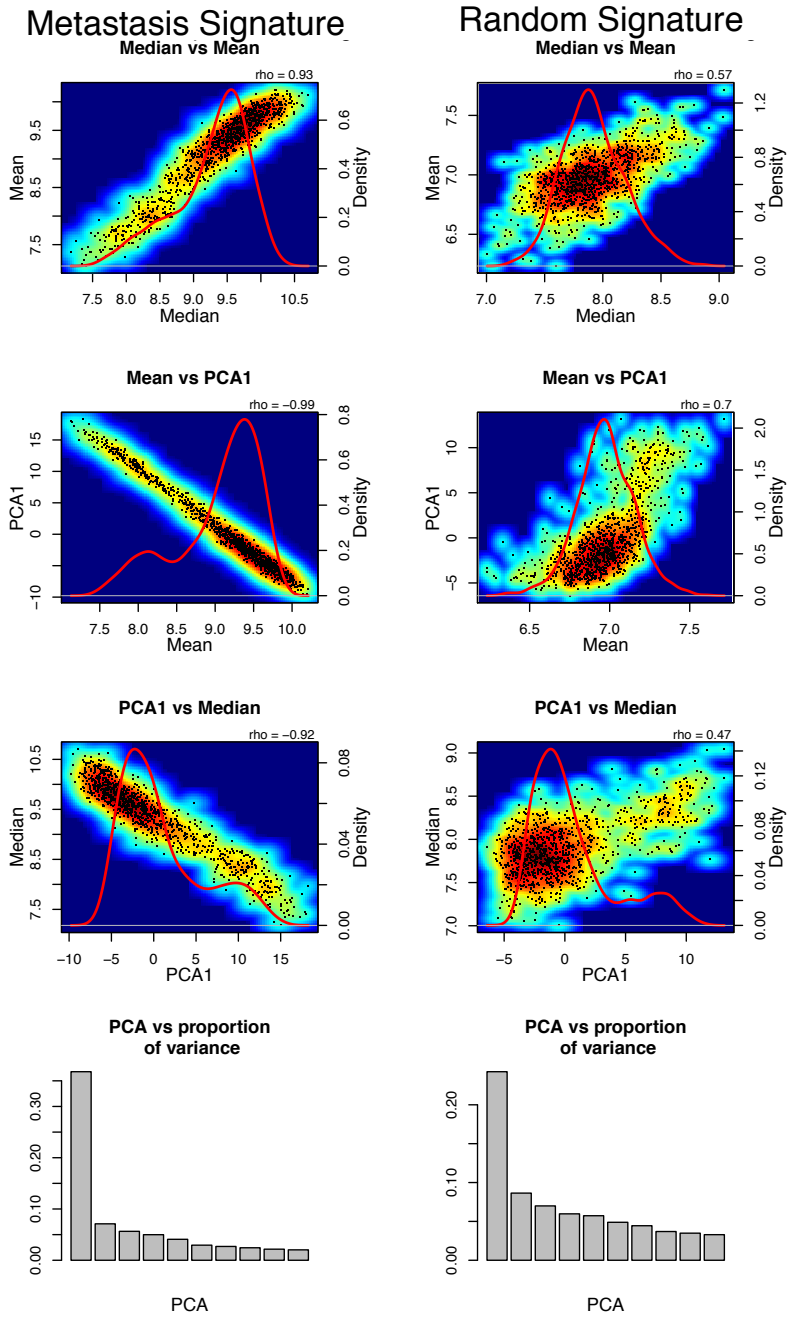

Figure 2.5: **Comparison of scoring metrics reveals metastasis signature may be summarised by mean, median, or first PCA component with little information loss.** Comparison of scoring metrics for a metastasis gene signature (left) and a set of random genes (right) in the TCGA breast cancer RNA-seq dataset.

all samples, in conjunction with annotations for the samples, if they were provided. These plots were clustered based on each dataset in turn, and run over each signature and each dataset present. An example of such a plot is shown in Figure 2.6, where the different expression profiles of the random gene signature and the metastasis gene

signature across patients can be observed. In addition to hierarchical clustering on patient samples and signature elements, biclustering of the signature genes' expression across patient samples was also attempted. Biclusters were not identified in either case of this analysis. In both cases of signatures that were considered, there were no clear subcomponents of signatures. If the random gene set was a true signature, these observations would suggest that the signature could be refined by removing the lowly expressed subset of genes.

Global comparison of multiple signatures: The plot in Figure 2.7 describes each signature applied to each dataset in a holistic, radar chart format. This plot evaluates the gene signature across a number of metrics, many of which are summary metrics for those in steps 4-9 of this procedure, and are described in detail in the Table 2.1. In this case, the radar plot summarises the stronger performance of the metastasis signature over the random gene set with respect to the quality control metrics considered.

Analysis of null distributions of quality control metrics: Lastly, to ascertain significance of the differences among metrics between both signatures in the radar plot, the null distributions were computed for each of the 14 metrics reported, for each signature and dataset combination. The values for each gene signature and dataset combination are shown in red overlaid with the other points in grey, giving a sense of significance of each of the metrics, as shown in Figure 2.8. From this figure, it may be observed that for the breast cancer metastasis signature on this dataset, the metrics evaluated show high significance for the signature genes, as compared to a random set of genes of the same length, whereas the same significance is not seen for the randomly chosen gene signature. In addition to these bootstrap resampling-based null distributions, analogous null distributions generated by permutation resampling may be presented, as described in the previous section of this chapter. This has been omitted for this case, as in this instance, the plot bears high similarity to the bootstrap resampling-based null distributions.

2.5 Example evaluation of signature translatability cross-platform

The second example considered is to compare the quality of the application of a signature on two datasets generated by different platforms. Signature transportability

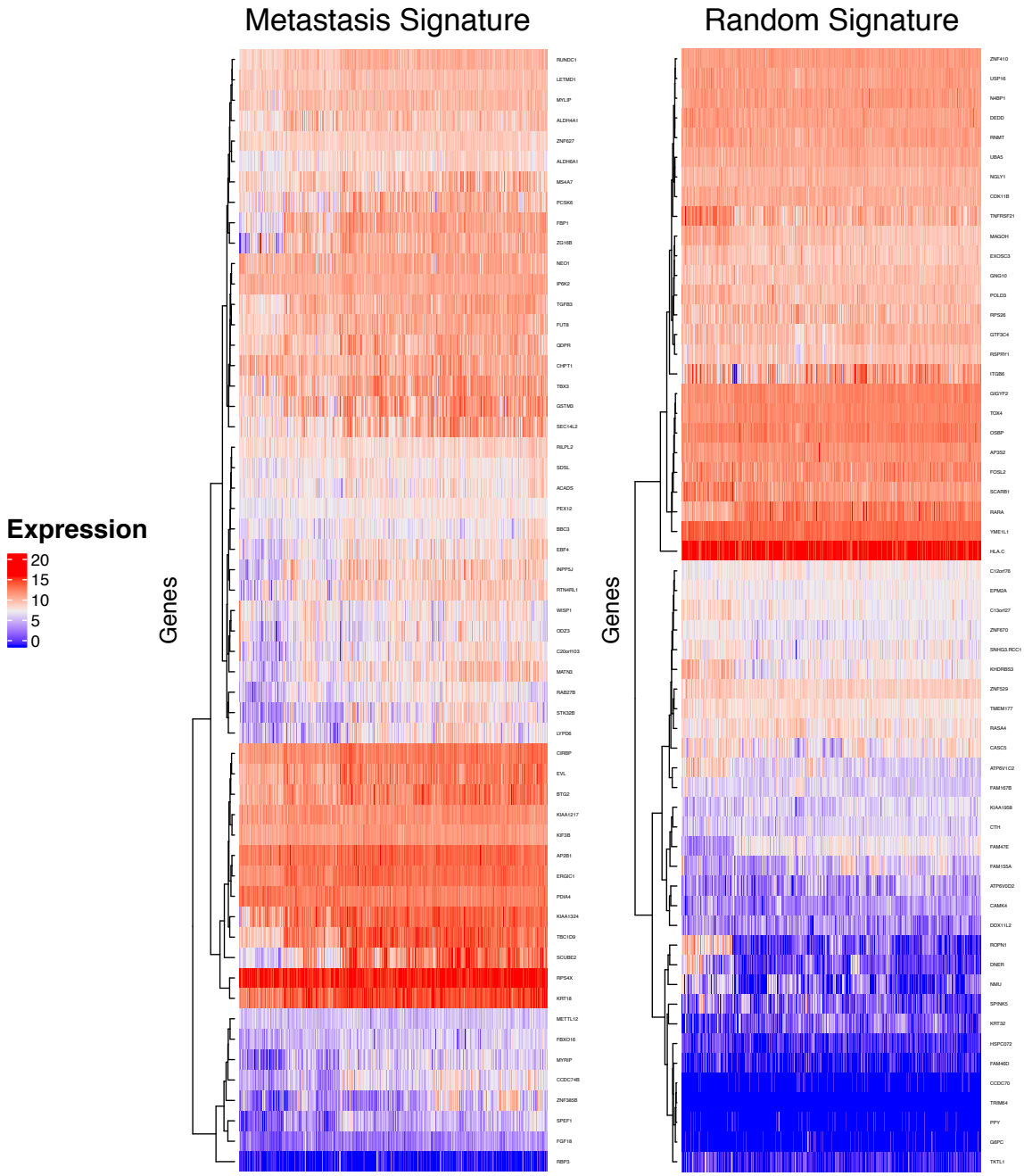

Figure 2.6: **Heatmaps show that metastasis gene signatures tend to be highly expressed across most samples.** Hierarchical clustering of signature gene expression for the metastasis signature (left) and the random gene set (right) over the TCGA breast cancer dataset.

is a major issue, as many gene signatures at present have been derived using microarray technology, whereas emerging datasets are primarily comprised of RNA-seq data. For this example, the breast cancer metastasis signature derived in [82], as used is the previous example, is reconsidered on each of an RNA-seq dataset (breast

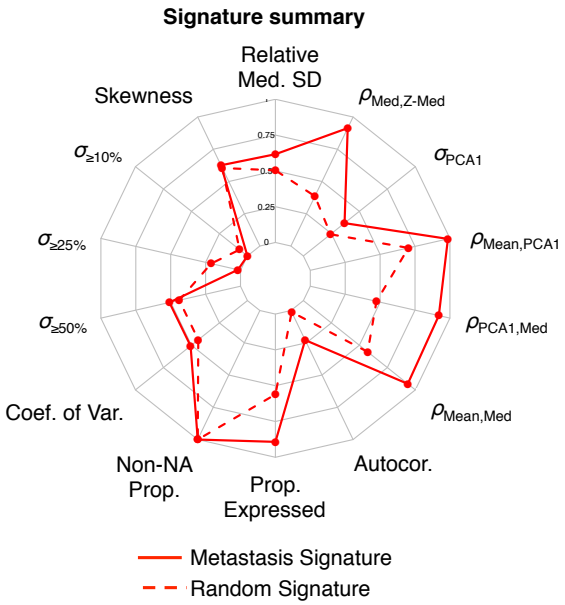

Figure 2.7: **Radar plot summarises differences between the signatures; the metastasis signature outperforms the random gene set on quality metrics.** Radar plot showing summary of gene signature quality control metrics for a metastasis signature (solid line) and a random set of genes (dashed line), as calculated for the TCGA breast cancer RNA-seq dataset.

cancer, TCGA, as was previously considered), and a microarray generated dataset (GEO Series GSE3494), as described above.

Signature annotation As an initial step, the Van't Veer metastasis signature was re-annotated to generate a list of probes compatible with Affymetrix U133A array technology, with probes mapping to signature genes, as well as possible. This pre-processing step was not included within *sigQC*, as many tools already exist that have this functionality, and these annotations change frequently with time; thus, tools receiving more consistent updates for these annotations, such as BioMart [84], are a better choice for this task.

Summary of quality control metrics For this example, the focus is applied not on the individual *sigQC* steps, but rather, the focus is on an analysis of the methodology from the radar plot produced at the end of the protocol. As shown in Figure 2.9, the metastasis gene signature behaves very well on both microarray and RNA-seq datasets, as evidenced by the overlap of the radar plot curves. It should be noted that, by their definition, the metrics represented on each of the axes of

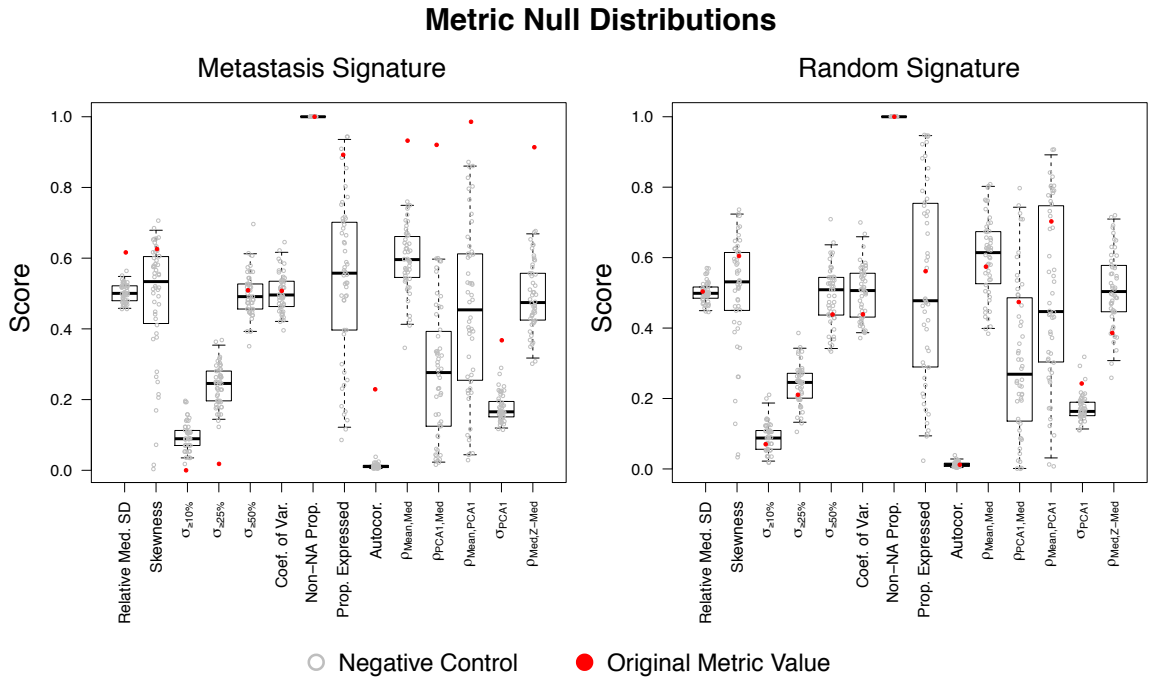

Figure 2.8: **The null distribution of each of the metrics reveals the significance of radar plot values for each signature.** Box and scatter plots depicting the null distributions of each of the metrics measured on the radar plot for the metastasis signature (left) and a random gene set (right), for $N = 50$ bootstrap resampling runs using random gene signatures of the same length.

the radar plot were normalised to the same interval, to facilitate a fair comparison between different datasets for the same signatures.

The bootstrap resampling-based null distributions of each of these metrics were next analysed, and these plots are depicted on the right side of Figure 2.9. Here it can again be observed that both signatures show strong performance on both datasets, with nearly all metric values highly statistically significant, as compared to random gene sets of the same length, suggesting that this is indeed a transportable signature between these two datasets. Moreover, it should be noted that when interpreting the null distributions of each of the fourteen radarplot metrics, this must be done within the context of the biology represented by the underlying dataset. For instance, the Van't Veer metastasis signature does not display a statistically significant proportion of genes with high variance or coefficient of variation, when compared to the null distribution in either dataset, but this is likely a quality attributable to the biology of the dataset itself. More specifically, because neither dataset contains non-tumour samples, the variability of all genes is reduced *a priori*, and this signature should not be interpreted as poorly performing on this dataset as a result.

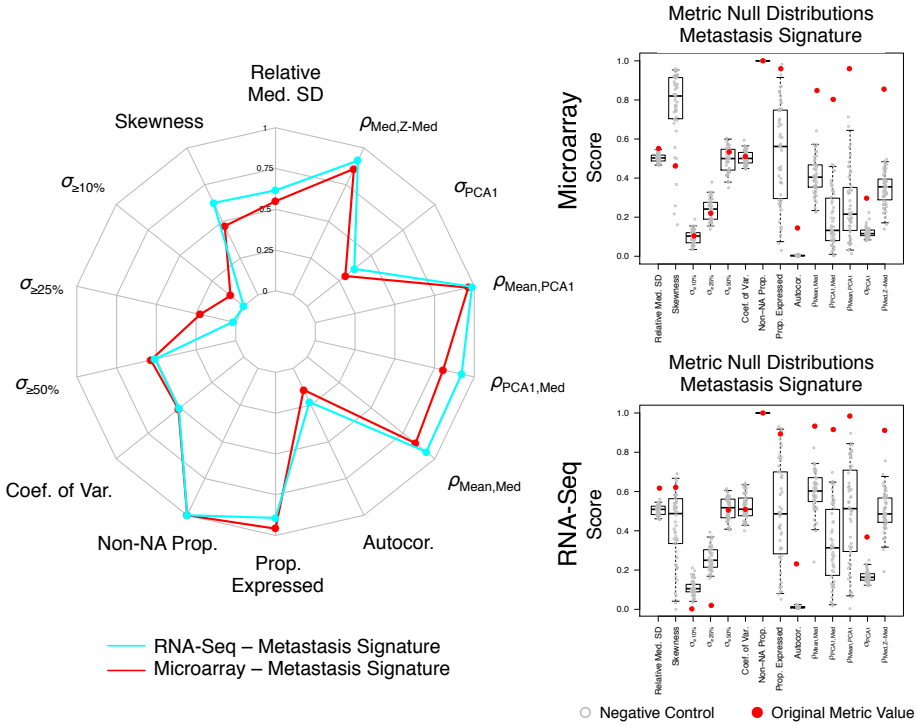

Figure 2.9: **Comparison of *sigQC* metrics for a metastasis gene signature on RNA-seq and microarray datasets.** Left: Radar plot summarising *sigQC* metrics for the metastasis gene signature considered on the RNA-seq (blue) and microarray (red) datasets for breast cancer samples. Right: Boxplots highlighting null distribution of each of the metrics considered, evaluated using $N = 100$ bootstrap resampling runs, for the metastasis signature on the microarray dataset (top) and RNA-seq dataset (bottom).

2.6 Discussion

2.6.1 Comparison with other methods of signature quality control

Prior to the design of *sigQC*, no generally adopted methods of gene signature quality control existed in the literature, though some methods were suggested for specific purposes. For example, a generally adopted technique for the validation of prognostic ability of a signature is to resample random gene lists of the same length as the gene signature, and determine their prognostic ability [76]. This resampling approach is among the techniques adopted by *sigQC* to provide the null distributions of the measured metrics. Such an analysis has also shown that a statistical significance cutoff of $p = 0.05$ may be too lenient when aiming to predict a specific clinical outcome for a gene signature, as many randomly selected gene signatures may also prognosticate at this level of statistical significance, suggesting a lack of specificity in predictive abil-

ity [76]. Furthermore, characteristics such as coherence, uniqueness, robustness, and transferability have been previously suggested in the context of signature evaluation [77], providing strong support for some of the *sigQC* metrics. However, their application has been limited to a small subset of gene signatures, namely those adopting PCA during the signature generation phase. Finally, consensus classification, addressing uncertainty arising from normalization and other data-processing steps, has also been proposed as an option to both evaluate and improve gene signature performance, but has not been widely adopted or implemented [93]. These methods solved existing issues related to gene signature development and validation, but primarily test specific aspects of the gene signature's performance, without enabling a broader assessment of its basic statistical properties across different contexts and datasets, nor an evaluation the qualities of the signature genes themselves.

2.6.2 Limitations

The protocol presented through *sigQC* is limited by the fact that the applicability of a gene signature to a broader setting can never be entirely determined, and so there may be characteristics, intrinsic to a signature or signature types, that enable it to pass all of the proposed quality control measures, without performing well in its intended sense. More succinctly, because *sigQC* does not account for the wide range of outcomes that gene signatures have been designed to predict, there is a limitation to a highly general solution which is not domain-specific. While *sigQC* provides metrics useful to the initial quality control for a gene signature and dataset, this does not optimise based on what is being predicted. For instance, *sigQC* does not take into account covariate adjustment when predicting specific outcomes, such as survival. Such a limitation will almost certainly occur, given the diversity of methodologies of gene signature generation and the wide variety of outcomes predicted, and to address this, it is cautioned that this protocol provides a set of conditions which are important to check, but are not fully sufficient for the determination of gene signature applicability. Undoubtedly, because of the nature of gene signatures, this limitation will be present in any broadly-scoped quality control methodology, as there may always be cases for which such a quality control methodology may not detect a poorly-performing signature.

A second limitation of *sigQC* relates to the number of signature scoring metrics possible and in use today. For simplicity and usability, *sigQC* supports basic primary summary metrics such as mean, median, and first principal component of the

expression of the signature genes. While these are strong, commonly used, and easily generalisable summary metrics, and they form the building blocks of many more complex metrics, others have been proposed that may show differences with what is used in *sigQC*. Alternative scoring metrics that have been proposed in the literature include many linear modelling approaches (for example, [94]), S-scoring [95, 96], averaged z-scores [97], and Pearson correlation based-scores [98]. The metrics employed by *sigQC* translate well to averaged z-scores, as a comparison of standardised data to unstandardised data is performed. Additionally, by testing intra-signature correlation using the Spearman correlation coefficient, it is possible to capture non-linear relationships not observed using Pearson coefficient-based signatures, at the cost of potentially increased noise from such non-linear relationships. S-scoring is based on a linear combination of z-scores, and combines the approaches of standardising the dataset with the directionality and flexibility of a linear model [95]. Thus, like a linear model, this scoring system, while it may be more flexible for defining dataset specific scores, often does not translate easily to new datasets or technologies. These methods are not tested explicitly in the current version of *sigQC*, however the metrics provided by *sigQC* provide a broad statistical assessment of the genes in a given signature across datasets and technologies; information which can be used to design more context-specific scoring techniques.

2.7 Summary and Conclusions

In this chapter a methodology was presented, by which gene expression signatures, defined as sets of genes displaying coherent expression patterns in conjunction with a biological process or clinical outcome, may be evaluated for their quality with respect to a given dataset. To standardise and simplify the computation of quality control metrics, an R package, called *sigQC* was developed, automating the computation of all of these metrics, available for download from CRAN.

Chapter 3

Pan-cancer characterisation of miRNA with hallmarks of cancer reveals negative association of miRNA expression with tumour suppressor genes

Abstract

miRNA are key regulators of the human transcriptome across a number of diverse biological processes, such as development, aging, and cancer, where particular miRNA have been identified as tumour suppressive and oncogenic. In this chapter, the association of miRNA expression and target regulation with the phenotypic hallmarks of cancer were elucidated across 15 cancer types comprising 7,316 clinical samples from TCGA. Utilising multivariable regression techniques to integrate transcriptomic, methylation and mutation data, statistically significant associations between miRNA expression and expression of genes related to the hallmarks of cancer were highlighted. This highlighted redundancy among the oncomiR-1 cluster of tumourigenic miRNAs, in particular hsa-miR-17-5p. In addition, exclusive statistically significant negative miRNA association was uncovered for the *PTEN*, *FAT4*, and *CDK12* tumour suppressor genes, potentially suggestive of an alternative mechanism of repression for these genes in the absence of mutation, methylation or copy number changes.

3.1 Introduction

3.1.1 The hallmarks of cancer define disease-associated traits

As discussed in the Introduction, the hallmarks of cancer outline the core phenotypic changes occurring in a tumour [4, 5]. These changes characterise cancer as a diseased state, different from normal tissue, and studying these differences may define actionable targets for therapeutic intervention. While it is argued that many of the traits individually may not be cancer specific, and the list remains incomplete, the combination of the hallmarks, especially invasiveness and metastasis, is highly relevant. Furthermore, from a clinical perspective, ultimately these traits are the key determinants of cancer-associated morbidity.

Despite the criticisms, since the definition of these characteristic hallmarks in 2001, and the subsequent genomic revolution that has occurred in the field of cancer biology, multiple groups have proposed gene expression signatures for these hallmarks [99, 100, 101]. These gene signatures generally consist of a set of tens to several hundred mRNA species, for which a summary metric of their collective expression associates with a known phenotypic hallmark, and may help with defining a therapeutic strategy [102]. Furthermore, since the second paper outlining the updated hallmarks by Hanahan and Weinberg was published in 2011, there has been a second revolution in the field of genomics; namely, the discovery of the diverse, critical roles of non-coding RNA in cancer.

As discussed earlier, previously thought to be junk DNA, non-coding RNA are those RNA derived from DNA that do not code for proteins, and consist of a diverse family of evolutionarily conserved species, including lncRNA, circRNA, and miRNA, among others [27, 28, 29]. Much effort has focused on the characterisation of these non-coding RNA, and early work has shown that these species, particularly miRNA, are involved in a number of cellular developmental, and differentiation processes [31]. miRNA exert their effects by complementary base-pair binding to a short 7-8 nucleotide seed region typically located on the 3' untranslated region of the mRNA which they inhibit [27, 103]. Whilst this complementary base-pair interaction defines many miRNA-target interactions, there is a class of non-canonical miRNA targeting that has been shown to occur throughout the transcriptome [104]. Such non-canonical interactions include many cases of imperfect seed matches, often with one base pair mismatch, and remain difficult to predict [104, 105, 106, 107, 108]. A single miRNA is thought to exert its repressive effects on hundreds to thousands

of transcripts, meaning that specific miRNA may have very wide-ranging effects on cellular phenotype [27, 109].

3.1.2 Phenotypically characterising miRNA is challenging

3.1.2.1 Detecting potential miRNA-mRNA repressive effects

Despite this potential, due to the highly variable effect on the single target transcripts and the many factors involved in post-transcriptional gene regulation in addition to miRNA, the repressive signal on their targets remains challenging to detect in clinical datasets, although this is being abridged by the availability of large genomic datasets, and has been shown through the TCGA [110, 111, 112, 113, 114]. These large-scale studies for miRNA-mRNA interactions in cancer have begun to leverage the power of clinical datasets with thousands of patients to detect small, context-specific effects [111, 112, 113, 114]. For instance, Jacobsen et al. studied the miRNA-target interactions recurring across cancer types in the TCGA datasets, and developed the CancerMiner web tool [115]. Through this work, Jacobsen et al. showed that multiple miRNA concurrently regulate the DNA demethylation machinery of the cancer cell, through effectors such as *TET1* and *TDG*, suggesting their important role in promoting cancer [115].

Moreover, *in vitro* miRNA mapping has raised the problem of reproducibility; as miRNA expression in cell lines has been shown to differ greatly from clinical specimens, and varies between experimental conditions [116]. This difficulty arises due to the highly variable effect on targeted mRNA transcripts and the many factors involved in post-transcriptional gene regulation in addition to miRNA. Thus, because the miRNA expression levels themselves are so variable, observing the repressive signal on predicted mRNA targets, as determined by sequence complementarity, remains even more challenging to detect in clinical datasets [110].

3.1.2.2 Variable expression of miRNA targets may impact miRNA function

A further complicating factor with respect to the study of miRNAs is the relative abundance of their predicted targets [117]. A given miRNA may have as many as thousands of predicted targets, with many experimentally verified, but often possessing large differences in function [118]. This has led to an almost paradoxical finding about the effects of miRNAs, in that a single miRNA may theoretically target mRNA

molecules with opposing effects in the cell [118]. This paradox is resolved by the observation that miRNA likely play different roles depending on the environment in which they are expressed [117, 119, 120]. Therefore, in addition to measuring the repressive effect of miRNA targets within a transcriptome, the effect of a miRNA on a transcriptome may vary massively, depending on the relative abundance of each of its targets. That is, a miRNA may only repress targets to which it is able to bind, and this requires the presence of the target in a detectable concentration relative to all others [121]. This means that the effect of a miRNA on phenotype can only reliably be understood for samples in which the expression of mRNA targets is comparable, underscoring how miRNA-mediated effects are highly context-dependent. Recent work has aimed at generating an understanding of how competing miRNA targets regulate each other, and work, in particular by Chiu et al. in [122] and Xu et al. in [123] has shown how these effects can be uncovered in a high-throughput manner.

3.1.3 Research questions

In this chapter, I show how elucidating potential miRNA-mRNA interactions can be done using gene signatures representative of the hallmarks of cancer. By classifying tumour transcriptomes using only the expression of genes associated to a particular phenotype with gene expression signatures, the statistically significant associations miRNAs with the hallmarks of cancer may be uncovered.

These results point towards a scenario wherein the transcriptome of the cancer cell, known to be driven by dysregulation of tumour suppressor genes and oncogenes, is heavily coassociated with miRNAs, extending the work by Jacobsen et al. and related studies by the TCGA consortium, and generating further experimental hypotheses [111, 112, 113, 114, 115]. In addition, it is shown that statistically significant miRNA-target associations predicted across multiple cancer types involve both tumour suppressors and oncogenes, potentially acting with oncogenesis. Study of these tumour suppressor genes yields novel conclusions about their regulation, particularly with respect to their repression by miRNA, in statistically significant association with their methylation, mutation or copy number alterations, and the exclusivity of the occurrence of these modes of regulation.

The remainder of this chapter is structured as follows. In Section 3.3.1, each of the gene signatures considered was checked for applicability and quality on the datasets to which they were applied, using *sigQC*, and then in Section 3.3.2 these mRNA gene expression signatures were used to develop a map of miRNA statistically significantly associated with the hallmarks of cancer, from which the miRNA families

involved were studied in Section 3.3.3.1. Next, in Sections 3.3.4 - 3.3.6, recurrent negative correlations between predicted targets of these mRNA were identified, with a specific focus on tumour suppressor genes, and it was shown that although the same miRNA were identified as signature-associated from normal tissues, these predicted target-miRNA correlations were not seen in the adjacent normal samples. Subsequently, in Section 3.3.8, the modes of regulation of these TSG were studied, through an integrated analysis of miRNA, mutation, copy number, and methylation. Then, in Section 3.3.9, an ad-hoc analysis designed to detect exclusivity in the miRNA association with other regulators of TSG expression was carried out, and it was shown how the different regulators of these TSG co-associate in expression across cancer types. Lastly, in Section 3.3.10, it was shown how the miRNA identified as negatively associated with tumour suppressor gene expression had potential clinical relevance in functioning as an independent classifier of the molecular subtype of breast tumours.

3.2 Methods

3.2.1 Gene signatures considered

A wide variety of gene signatures was considered, corresponding to many of the hallmarks of cancer, as described in the original and updated work by Hanahan and Weinberg [4, 5]. Signatures were selected through a review of MSigDB hallmarks signatures, as well as through a review of the literature, and are summarised in Table 3.3 [69]. While many of these signatures were derived for a particular tumour type, these have been applied across multiple tumour types, but before doing so, an evaluation step (*sigQC*) was performed, to ensure that each signature used was applicable to every dataset under consideration, as described in Appendix B1, Figures B.1- B.9.

3.2.2 Datasets considered

In selecting datasets for this analysis, those comprising a comprehensive set of cancer types were sought, with each tumour type represented by a sufficient number of clinical samples, so as to reduce the effects of noise.

TCGA Initial consideration was given to all cancer types represented within the TCGA dataset, and limited based on origin of neoplasm and number of patients for whom miRNA-sequencing was carried out [25]. The RSEM normalised gene expression, mature miRNA normalised expression data, copy number, mutation, and methylation data were accessed from the Firebrowse database at http://www.firebrowse.org,

accessed on November 11, 2017. In particular, all cancer types which were epithelial carcinomas or adenocarcinomas histologically, and with at least 200 samples with miRNA-sequencing data were included for analysis. These two filters limited the cancers considered to a total of 15 epithelial or glandular neoplasms, including 7,738 clinical samples, of which 7,316 had miRNA-sequencing data. The tumour types, along with their sample counts are presented in Table 3.1. Details of the number of samples included for each data type are presented in Table 3.2. Any dataset present with fewer than 9 samples was excluded from analysis. This restriction excluded the analysis of COAD, OV, and UCEC datasets from the analysis of tumour suppressor genes, oncogenes, and exclusivity of regulation.

Metabric The Metabric dataset is also used in this section, as an additional external dataset for which the reproducibility of miRNA-signature statistically significant associations could be tested. This dataset consists of 2136 fresh frozen primary breast tumour specimens, with associated clinical annotation, collected from tumour banks in the UK and Canada [124]. Molecular analysis of the samples with regard to transcriptomic profiling, was carried out for UK samples by first extracting DNA and RNA from ten 30μm sections from fresh frozen tumours using the Qiagen DNeasy Blood and tissue kit and the Qiagen miRNeasy Kit [124]. For Canadian samples, DNA and RNA were extracted from ten 8μm sections from fresh frozen tumours using the MagAttract DNA Mini M48 Kit and Qiagen miRNeasy 96 kit [124]. Using these specimens, following quality control, RNA were prepared using the Illumina Totalprep RNA Amplification kit and then RNA were hybridised to the Illumina HT-12 v3 Expression Beadchips per manufacturer instructions [124]. Raw expression data was preprocessed using a custom R script, and low quality samples were removed [124]. Intensity data was then quantile-normalised, and linear modelling was used to correct for any batch effects related to positioning of arrays on the Illumina BeadChip [124]. These normalised data were then log2 scaled, and these represent the data used for all further analysis as presented.

OV-AU An independent dataset as part of the International Cancer Genome Consortium (ICGC) dataset was considered for further evaluation of the presented results in samples of ovarian serious cystadenocarcinoma. More specifically, the OV-AU project from the ICGC data portal was comprised of paired miRNA, mRNA expression, methylation, copy number, and mutational data for 93 samples [125]. Samples were collected and sequenced in Australia, from women diagnosed with epithelial,

Dataset	Abbreviation	Clinical samples
Breast invasive carcinoma	BRCA	1098
Ovarian serous cystadenocarcinoma	OV	602
Lung adenocarcinoma	LUAD	585
Uterine corpus endometrial carcinoma	UCEC	560
Kidney renal clear cell carcinoma	KIRC	537
Head and neck squamous cell carcinoma	HNSC	528
Lung squamous cell carcinoma	LUSC	504
Thyroid carcinoma	THCA	503
Prostate adenocarcinoma	PRAD	499
Colon adenocarcinoma	COAD	460
Stomach adenocarcinoma	STAD	443
Bladder urothelial carcinoma	BLCA	412
Liver hepatocellular carcinoma	LIHC	377
Kidney renal papillary cell carcinoma	KIRP	323
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	307

Table 3.1: **TCGA datasets considered and associated total clinical sample counts.**

primary peritoneal, or fallopian tube cancer from 1992-2012. Samples were frozen and were required to have at least 70% tumour cells on histologic staining. Those passing this criteria underwent nucleic acid extraction via the DNeasy blood and tissue kit (Qiagen), and RNA isolated with the mirVana miRNA Isolation kit (Ambion). SNP and copy number was assayed with the Omni 2.5-8, v1.0 and v1.1 IlluminaBeadChips. RNA-seq libraries were prepared using the TruSeq RNA Sample Preparation v2 kit (Illumina), and were paired-end sequenced on the HiSeq2000 platform. Methylation was assessed using the Infinium Human Methylation 450 Bead Chip (Illumina). miRNA levels were determined using the Nanostring nCounterHuman v2.1 miRNA expression analysis kit.

3.2.3 miRNA family database

miRNA ranked across different cancer types were further grouped together by miRNA family, as defined by the targetscan database, implemented in R as the targetscan.Hs.eg.db package [126, 127].

Dataset	mRNA samples	miRNA	mRNA and miRNA	mRNA, miRNA, mutation, methylation, and copy number
BRCA	782	755	499	324
OV	307	461	291	0
LUAD	517	452	449	181
UCEC	177	412	174	4
KIRC	534	255	255	121
HNSC	520	486	478	244
LUSC	501	342	342	51
THCA	501	502	500	396
PRAD	497	494	493	329
COAD	286	221	221	0
STAD	415	389	370	230
BLCA	408	409	405	128
LIHC	373	374	369	186
KIRP	291	292	291	148
CESC	304	307	304	190

Table 3.2: **Counts of common samples with miRNA, mRNA, mutation, methylation, and copy number data.**

Signature name	Reference	Genes (#)	Associated hallmarks
Epithelial Mesenchymal Transition, MSigDB	MSigDB [69]	200	Activating invasion, metastasis
Invasiveness	Marsan et al., 2014 [128]	16	Activating invasion, metastasis
Oxidative Phosphorylation, MSigDB	MSigDB [69]	200	Deregulating energetics
Reactive Oxygen Species Pathway, MSigDB	MSigDB [69]	49	Deregulating energetics
G2M Checkpoint, MSigDB	MSigDB [69]	200	Enabling immortality
PI3K-AKT-MTOR Signaling, MSigDB	MSigDB [69]	105	Evading growth suppressors
Xenobiotic Metabolism, MSigDB	MSigDB [69]	200	Evading growth suppressors
DNA Repair, MSigDB	MSigDB [69]	150	Genome instability, Enabling immortality
p53 Pathway, MSigDB	MSigDB [69]	200	Genome instability, Enabling immortality
Hypoxia	Buffa et al., 2010 [74]	51	Inducing angiogenesis
Angiogenesis, MSigDB	MSigDB [69]	36	Inducing angiogenesis
Hypoxia, MSigDB	MSigDB [69]	200	Inducing angiogenesis
Angiogenesis, upregulated	Desmedt et al., 2008 [21]	5	Inducing angiogenesis
Angiogenesis	Masiero et al., 2013 [129]	43	Inducing angiogenesis
Apoptosis, MSigDB	MSigDB [69]	161	Enabling replicative immortality
Apoptosis	Desmedt et al., 2008 [21]	4	Enabling replicative immortality
Proliferation, upregulated	Desmedt et al., 2008 [21]	140	Sustaining proliferative signaling
KRAS Signaling, Up, MSigDB	MSigDB [69]	200	Sustaining proliferative signaling
Inflammatory Response, MSigDB	MSigDB [69]	200	Inflammation, Avoiding immune destruction
IL2-STAT5 Signaling, MSigDB	MSigDB [69]	200	Inflammation, Avoiding immune destruction
IL6-JAK-STAT3 Signaling, MSigDB	MSigDB [69]	87	Inflammation, Avoiding immune destruction
TGFβ Signaling, MSigDB	MSigDB [69]	54	Inflammation, Avoiding immune destruction
TNFα Signaling via NF-κB, MSigDB	MSigDB [69]	200	Inflammation, Avoiding immune destruction
Immune Invasion, upregulated	Desmedt et al., 2008 [21]	92	Inflammation, Avoiding immune destruction

Table 3.3: **Gene signatures considered and associated hallmarks of cancer.**

3.2.4 Statistical methodology

3.2.4.1 Transcriptomic data

Data were taken from the Broad Institute genomic data access centre (GDAC) Firebrowse TCGA portal. miRNA datasets used were log2 normalised mature miRNA counts for all cancer types. mRNA datasets used were normalised RSEM genes taken from data through the Illumina HiSeq RNA-seq v2 platform. These expression data were then transformed by the transformation $\log_2(x+1)$, for x as the original expression value, and this was used in all further computation for all cancer types and signatures. Where not otherwise specified, signature scores were taken as the median of log2-transformed expression of all signature genes for each sample. Metabric datasets for normalised miRNA and mRNA expression were taken from the European Genome-Phenome Archive (EGA) under study accession numbers EGAD00010000434 and EGAD00010000438. In all analyses, only miRNA and mRNA expressed at a non-zero level in at least 80% of samples were considered.

3.2.4.2 Penalised linear regression

The aim of the penalised linear regression methodology was to determine those miRNA which most strongly predict (positively or negatively), the gene expression summary score for each signature. With consideration of this, the linear regression was designed such that the model utilised the expression levels of each individual miRNA as a covariate, in order to predict the gene signature score. Gene expression signature scores were taken as the median value of the signature genes' expression, where mRNA gene expression was used as the log-transformed RSEM value as defined above. Likewise, the miRNA expression used was log transformed as well, as described above. Within the statistical model, to facilitate more direct comparability between coefficients, gene expression signature scores and miRNA expression levels were scaled to a mean of zero and unit variance using the z transform.

A statistical approach involving a combined univariate-multivariate penalised linear regression was applied, with 10-fold cross validation to infer statistically significant relationships between miRNA and gene signatures, without overfitting the model. A combined univariate-multivariate approach was chosen in this instance to aid with feature selection. Because the total number of samples per cancer type was generally less than 500, the number of predictors (initially 2500 miRNAs) had to be reduced for appropriate statistical modelling. To accomplish this, an initial univariate filter was applied to remove miRNA showing little predictive power from the multivariate

linear model, and only those miRNA with $p < 0.2$ significance in the univariate linear model predicting signature score were considered. This permissive p-value was used as a balance to both remove miRNA predictors, but also to ensure that the multivariate linear model did not contain artificially stringent statistically significant associations, as the penalisation procedure also functioned as a stringency filter, thereby reducing the false discovery rate. Following this, with this reduced set of predictors, the linear regression could be carried out. Elastic net, or combined L1/L2 penalised linear regression was chosen for this, for two reasons. First, the L1 component of the penalty term in the elastic net acted to increase interpretability of results by pushing coefficients towards zero, and thereby reducing the number of non-zero coefficients, and a compromise was obtained with this to ensure it was not too severe of a penalty with the L2 term. In effect, the L2 term mitigated the effect of the L1 term reducing coefficient values to zero by reducing the value of coefficients in the model in tandem (as opposed to reducing one to zero), and also reducing coefficients in tandem for terms that were correlated between themselves. Using this hybrid approach, with a range of ratios between L1 and L2 penalties, and then considering the model with the best fit, as measured through the log likelihood, allowed for the benefits of each approach to be used, which were particularly valuable for this high dimensional dataset.

To tune the parameters for the combined L1/L2 regression, a range of values (0, 0.01, 0.1, 1, 10, 100), was tested for the L2 parameter, while in each case the L1 parameter was optimised. Following computation of all models, the model with the greatest log-likelihood was chosen. All model-fitting was done with 10-fold cross-validation, and was carried out using the *penalized* package in R, version 0.9-50 [130, 131]. This model was fit independently for each tumour type, and then miRNA coefficients were then later aggregated and ranked by value across tumour types. Those miRNA that had non-zero coefficients across at least 5 tumour types, and that were recurrently highly ranking compared to all others, across tumour types, were selected by the rank product test, with statistically significant miRNA defined as those with Bonferroni-corrected p value less than 0.05. The same approach was used to identify miRNA with consistently negative coefficients across cancer types.

3.2.4.3 Rank product analysis

Once coefficients were obtained for the linear model via the penalised regression approach described earlier, these were collated into matrices with columns defined by cancer type, for each of the gene signatures considered. These coefficients were then fractionally-ranked both from most negative to most positive, and most positive to

most negative in value. The rank product statistic, as described by Breitling et al., in 2004, for these fractional ranks, was then considered, and the coefficients were ranked in terms of their significance of rank product test statistic, as implemented by the RankProd R package, version 3.6.0 [132, 133], by Bonferroni-corrected p values and a corrected significance threshold of $p < 0.05$. This was used to give rankings of miRNA associated both positively and negatively with the various signatures considered.

3.2.5 Analysis of predicted targets

Targets were aggregated for each miRNA using the miRNAtap database in R, version 1.14.0, as implemented through the Bioconductor targetscan.Hs.eg.db package, version 0.6.1 [134]. The miRNAtap R package is a resource that can be queried to obtain an aggregated list of predicted mRNA targets for a given miRNA. In this package, the default settings of using all 5 possible target databases: DIANA version 5.0 [135], Miranda 2010 release [136], PicTar 2005 release [137], TargetScan 7.1 [138], and miRDB 5.0 [139], with a minimum source number of 2 were used, and the union of all targets found was taken as the set of targets for a given miRNA.

For each of these predicted target-miRNA pairs, the Spearman correlation coefficient was calculated across every cancer type for miRNA versus target mRNA expression, partial to mutation status of the mRNA, and if this value reached statistical significance of $p < 0.05$, it was recorded; otherwise it was recorded as 0. Note that mutational status was reported as a binary variable with a value of 1 for any non-silent, non-intronic mutation, and 0 otherwise. The target-miRNA pairs with at least 5 non-zero entries across cancer types were kept for further analysis, and subsequently were analysed using the rank product statistic, to identify those pairs with consistently negative correlations, across cancer types, with respect to all other hallmarks-miRNA pairs. Partial correlations were done in R using the ppcor package, version 1.1 [140].

Furthermore, in the global analysis of all TSG-miRNA pairs, TSG-miRNA predicted target pair was considered, and the Spearman correlation partial to mutation status was also considered, collapsing the value to 0 if significance $p \geq 0.05$. The rank product statistic was again considered on those pairs with at least 5 non-zero values across cancer types, thereby identifying those TSG-miRNA pairs consistently negatively correlated across cancer types, statistically significant with respect to all other TSG. Lists of known oncogenes and tumour suppressor genes were taken from the COSMIC database [141]. Because *MYC* amplification was a possible confounder to the miRNA identified as associated with TSG across cancer types, the association

of *MYC* amplification and TSG mutation was checked across cancer types, and this was not found to co-occur or show statistical significance. Of the 96 TSG-cancer type pairs (8 TSG over 12 cancer types), none showed significance in the over-enrichment by a one-sided Fisher exact test for *MYC* amplification and TSG mutation after correcting for multiple testing.

3.2.6 Analysis of TSG regulation

In analysing the potential regulation of the TSG identified as related to the hallmarks of cancer and potentially amenable to miRNA regulation, the samples under consideration were first limited to those where copy number data, gene expression data, miRNA expression, mutation data, and methylation data were all present. Mutation data was taken as a binary variable, such that mutations were stratified into their reported types (e.g. missense mutations were all grouped together, etc.). For example, the missense mutation variable only contained a value of 1 if the sample had a missense mutation in the gene of interest, and was 0 otherwise. All variables considered in the linear regression were standardised to a mean of 0, and a standard deviation of 1. Methylation data considered was comprised of beta values, representing the fraction of methylated probe sites across a given sample for the probe sets present on the Illumina 450K methylation array. For each TSG, the probe sets considered were those annotated by the Illumina documentation with location on known promoter sites for the TSG and those within the coding sequence of the TSG itself.

L1/2 penalty-based penalised linear regression was then performed, in the same manner as above, for the linear model described in Figure 3.6a. Subsequently, coefficients were aggregated across the various cancer types and after the rank product test was applied, those predictors showing statistically significant consistent positive or negative coefficients were identified. Following this, the autocorrelation of each of these predictor variables was considered, for each of the TSG in each cancer type, as depicted by the heatmap in Figure 3.7a.

3.2.7 Analysis of the exclusivity of putative negative regulators associated to TSG

To determine the exclusivity of the co-expression of the putative negative regulators for each TSG, the empiric distributions of the variables Π_{ρ_k}, as defined graphically in Figure 3.7, were calculated. These represent the proportion of miRNA-miRNA or miRNA-methylation or methylation probe-methylation probe pairs that showed

statistically significant positive Spearman co-correlation ($p < 0.05$). For the bootstrapping analysis, the datasets were resampled such that miRNA and methylation probes were chosen in the same number as the heatmap in question, and then distributions for the pairwise differences in the variables Π_{ρ_k} were computed. From these distributions for the pairwise differences, the percentile on the empirically constructed cumulative distribution function (CDF) that the observed case represented could be inferred; the results of which are depicted in Figure 3.7b. This figure shows, for each gene and cancer type, the percentile on the pairwise difference empiric distribution for the variables Π_{ρ_k} defined from the observed heatmap.

3.3 Results

3.3.1 Evaluation of hallmark gene signatures across cancers

Gene signatures as listed in Table 3.3, were chosen to be representative broadly of the hallmarks of cancer, and were applied across datasets from multiple tissue types taken from the TCGA. *sigQC* version 0.1.20 was applied on all combinations of 15 datasets and 24 signatures considered, and this tested the consistency of signature performance across cancer types, which gave confidence in the application of the signatures to these datasets [142]. As assessed by the area contained within the radar plot for each dataset-signature combination, all areas were within 15-50% of the total possible area, suggesting that performance was relatively consistent. All summary plots from the *sigQC* quality control protocol are presented in Appendix B2. Each of the signatures considered over the 15 epithelial cancer datasets showed good applicability specifically with, strong signature gene expression (no NA expression, and at least 65% of all signature genes expressed in all but two cases), moderate gene signature score variability (coefficient of variation above 25th percentile in all cases), and moderate-strong autocorrelation of signature metrics (above $\rho = 0.8$ in approximately 75% of cases). These findings justified the subsequent use of these signatures across this pan-cancer dataset, to identify conserved statistically significant associations of miRNA and signature gene expression across epithelial tumours.

3.3.2 Hallmark gene signatures association analysis reveals a complex, statistically significant, pan-cancer miRNA association network

To determine the statistically significant association of gene signatures to miRNA expression, the signature score (median expression of signature genes) was set to equal a linear model consisting of all miRNAs that showed at least moderate univariate predictive ability (those with p-value ≥ 0.2 in an analogous univariate linear model were removed) for the signature summary score, as depicted in Figure 3.1a. Multivariable linear modelling with L1/L2 penalized regression optimized by ten-fold cross-validation was used to identify the miRNAs which showed the greatest predictive ability for each hallmark signature score, in each cancer type individually. An example of the values for miRNA coefficients across cancer types following the model fitting is depicted in Figure 3.1b. miRNAs were then ranked based on their final model coefficient (reflective of the strength of statistically significant association to the signature score), and miRNAs which ranked consistently high as positive predictors for a given hallmark signature score across cancer types were aggregated, from which statistically significant miRNAs were isolated with the rank product test (signature-associated miRNAs). Likewise, for each gene signature, the miRNAs most consistently ranked as strong negative predictors of signature score across cancer types were aggregated by the same rank product-based methodology (negatively signature-associated miRNA), as depicted in Figure 3.1c. This analysis revealed both many known and unknown statistically significant associations for miRNA and gene signature scores.

As an example highlighting the validity of many of these predictions, the case of the miRNA found to associate statistically significantly with the hypoxia signatures was considered. Hypoxia is one of the most studied microenvironmental perturbations in the context of miRNA regulation, and one with a very well-defined pathway, controlled largely by a single transcription factor protein, HIF-1α [143]. The intersection of the sets of miRNAs found to associate positively with the two hypoxia gene signatures (Hypoxia, Buffa et al. [74], and Hypoxia, MSigDB [69]) was taken, and this yielded predictions for hypoxia-associated miRNAs, across tumour types.

As shown in the tables in files associated with Appendix B3, this analysis identified many miRNAs previously associated with both hypoxia gene signatures, including hsa-miR-210-3p ($p < 10^{-18}$) [144], -21-3p ($p < 10^{-9}$), -21-5p ($p < 10^{-4}$), -23a-5p ($p < 10^{-4}$), -23a-3p ($p < 0.002$), -24-3p ($p < 10^{-5}$), -24-2-5p ($p < 10^{-3}$), -27a-5p ($p < 10^{-4}$), [145], let-7e-5p ($p < 10^{-4}$), let-7e-3p ($p < 10^{-3}$) [146], -22-5p ($p < 0.01$),

-22-3p ($p < 0.01$) [147]. This analysis also suggested statistically significant, pan-cancer associations for other members of the let-7 family of miRNAs in hypoxia; namely, let-7b-5p ($p < 10^{-5}$), let-7b-3p ($p < 0.002$), let-7d-5p ($p < 0.004$), let-7d-3p ($p < 0.004$), as well as hsa-miR-223-3p ($p < 10^{-6}$), -18a-5p ($p < 0.005$), and -28-3p ($p < 0.004$), which were previously unidentified. Note that the p values listed here are for the rank product statistic for each of these miRNA, across both signatures, where the maximum p value was typically taken, for those miRNA in common; these were only those miRNA in common to both signatures showing positive, statistically significant, association with each of the signatures, with significance determined by the rank product statistic, with Bonferroni correction for multiple testing.

From the gene signatures considered, a map connecting each miRNA to each signature with statistically significantly associated hallmarks was created. As shown in Figure 3.1d, this is a highly interconnected and complex network, with a core set of miRNAs shared across the hallmarks of cancer. A similar analysis produced an analogous map for the miRNA-hallmarks network for the miRNA negatively associated with both signatures, as shown in Figure 3.1e. To test the reproducibility of these results, the signature-miRNA linear model was rebuilt using the independent Metabric cohort, with the same methodology [124].

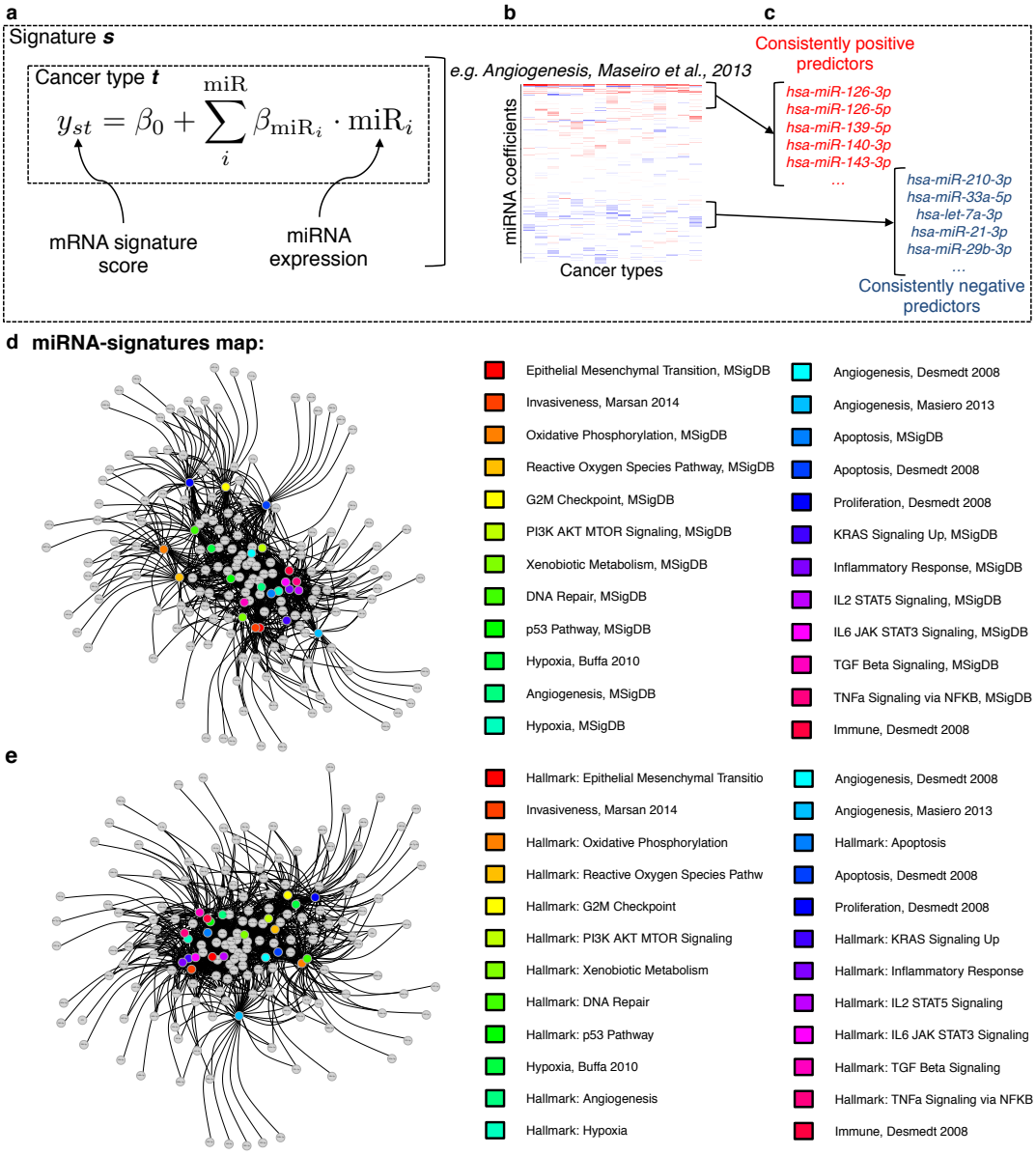

Figure 3.1: **Overview of approach used to identify hallmarks-associated miRNA.** (**a**) Overview of the linear model used in the fitting, for each gene signature and cancer type under consideration. (**b**) Example of a heatmap depicting the values of the coefficients identified for the miRNA predictors (rows), across cancer types (columns) for the Masiero angiogenesis signature [129]. (**c**) Consistently positive and negatively ranking miRNA coefficients, identified as statistically significant by the rank product statistic, are taken as the positive and negative hallmark-associated miRNA for a given hallmark signature. (**d**) Network map of signatures (coloured nodes) and their positively associated miRNA (grey nodes), connected by edges when a statistically significant association was found, highlighting strong interconnectivity between distinct molecular signatures. (**e**) Network map of signatures (coloured nodes) and their negatively statistically significantly associated miRNA (grey nodes), connected by edges when a statistically significant association was found, highlighting strong interconnectivity between distinct molecular signatures.

3.3.3 Reproducibility of miRNA-signature statistically significant associations

The reproducibility of the results for the miRNA-signature network is depicted in Figure 3.2. There was statistically significant overlap between the miRNA identified as positively or negatively associated to each signature among both datasets, by way of two-sided Fisher's exact test. In order to ensure reproducibility of the approach used to identify gene signature-associated miRNA, the linear modelling procedure was repeated across the independent Metabric matched miRNA and mRNA microarray dataset of 1293 samples [124]. Each gene signature was mapped to corresponding Ensembl IDs, and repeated the combined univariate-multivariate linear modelling approach was used over all miRNA probes. The miRNA probes with positive and negative coefficients were then identified, and mapped to their corresponding mature miRNA ID. Nearly all signatures showed statistical significance, and in the majority of cases not reaching statistical significance, signature applicability to the Metabric dataset presented a small issue, as signatures contained a high proportion of genes with low variance, which presents an issue for signature applicability, particularly for microarray-based datasets.

3.3.3.1 miRNA families associated with hallmarks signatures may possess both tumour suppressive and oncogenic roles

Subsets of miRNAs that typically share common, evolutionarily-conserved sequences or functional motifs in their mature or immature sequences are grouped into families [148, 149]. However, grouping the miRNAs found to be up- and down-regulated in statistically significant association with each of the gene signatures, as defined through the derived miRNA network, revealed that a number of miRNAs from the same families were present in both up- and down-regulated groups. That is, as summarised in Figure 3.3, there were cases where the same miRNA families contained miRNAs both positively and negatively statistically significantly associated with gene signatures for the hallmarks of cancer. In particular, the miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519 and let-7/98/4458/4500 families had multiple members across signatures both in statistically significant positive and negative associations. It should be noted, however, that these statistically significant discordant associations with miRNA family members across the gene signatures for the hallmarks of cancer may also have arisen as a result of spurious associations, due to noise

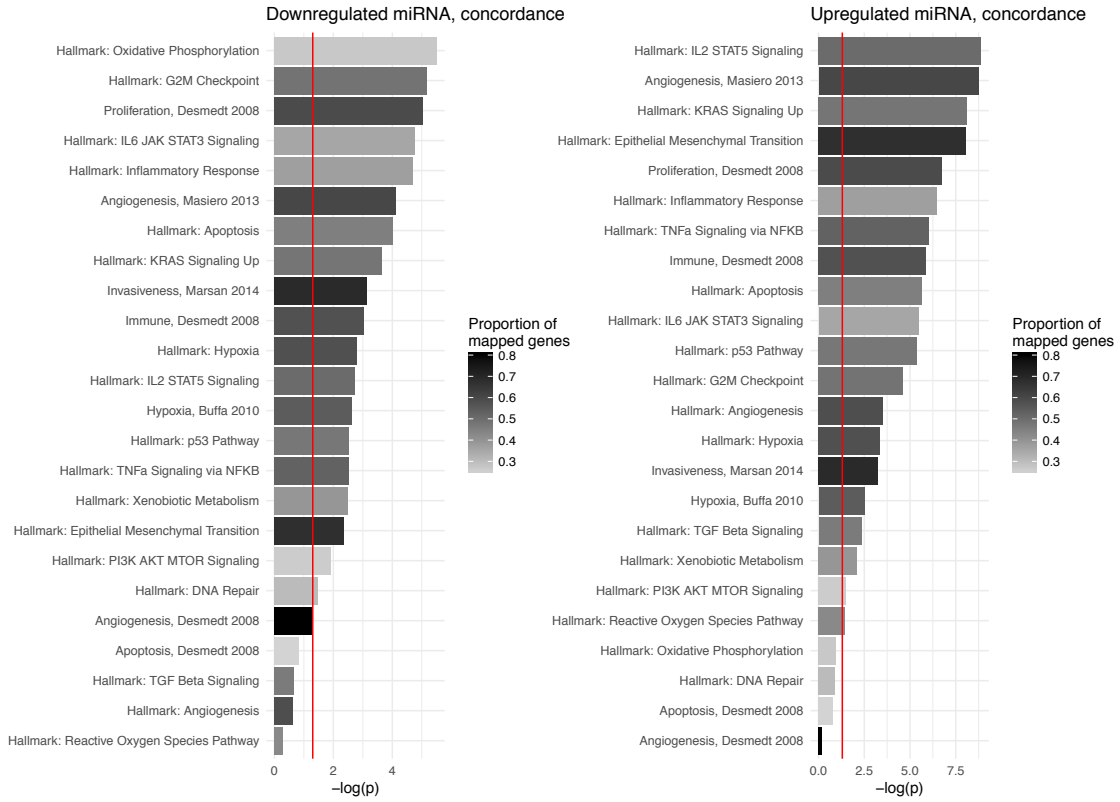

Figure 3.2: **Results of statistically significantly signature associated miRNA validation on the Metabric dataset.** Bar chart depicts -log of the p-value for the two sided Fisher exact test examining the overlap of the statistically significantly signature associated miRNA identified as positive (right) or negative (left) predictors of gene signature score in the Metabric breast cancer dataset as compared to those identified by an analogous approach in the TCGA pan-cancer dataset. Vertical red bar highlights the significance cutoff of $p = 0.05$. Bars are coloured as a function of proportion of signature genes mapped to mRNA probes present within the Metabric dataset.

in the dataset or a lack of discriminatory statistical power. As such, these divergent associations must first be further substantiated by more data and potentially experimental evidence, before any definitive conclusions can be made.

3.3.4 Predicted hallmarks-associated miRNA targets are statistically significantly enriched for tumour suppressor genes

A list of positively associated miRNA to the cancer hallmarks was first defined, specifically comprised of those miRNA that showed statistically significant association with at least one gene signature considered, defined as a Bonferroni-corrected p value less

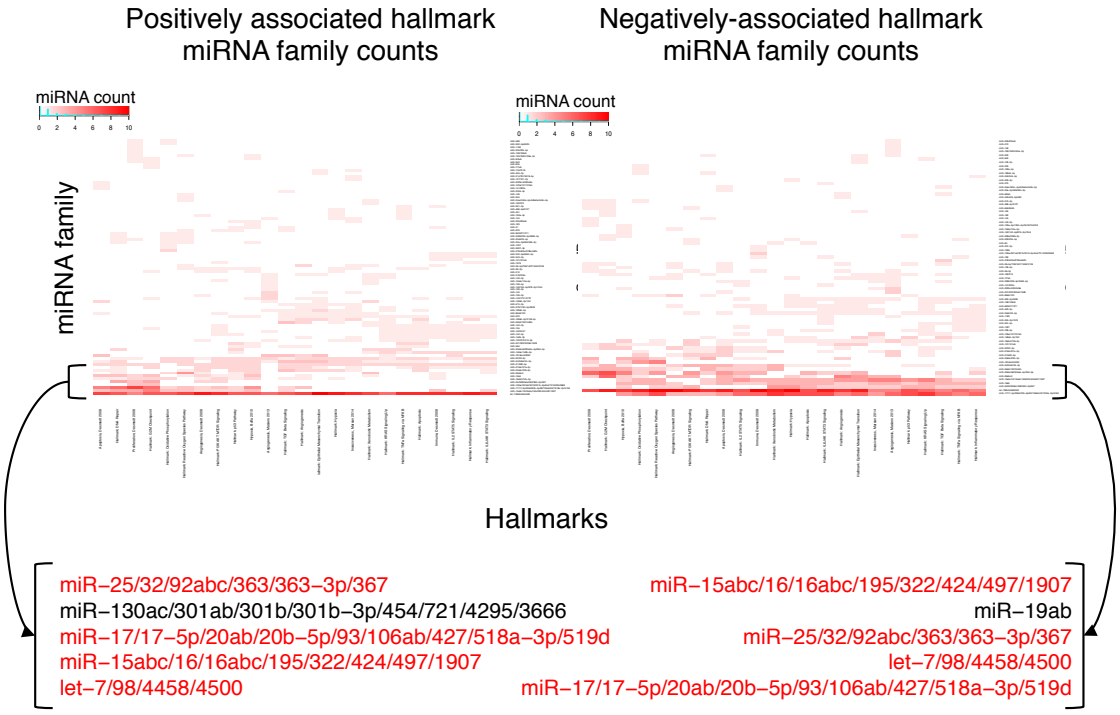

Figure 3.3: **miRNA families contain statistically significantly associated miRNA both positively and negatively with hallmarks gene signatures.** Heatmaps of counts of miRNA positively (left) and negatively (right) associated with each miRNA family (rows), aggregated by gene signatures (columns). Top 5 rows from both heatmaps are enlarged, with common entries highlighted in red, showing strong concordance between both the up- and down-regulated miRNA families across signatures and cancer types.

than 0.05 in the rank product test for Spearman correlation of miRNA expression to gene signature score across the 15 cancer types considered. Using these positively associated miRNA, predicted miRNA-target pairs were analysed, and those that showed statistically significant negative correlative association across multiple cancer types were identified. Targets were predicted using the union of five miRNA target prediction algorithms, as implemented by the Bioconductor package miRNAtap version 1.14.0 [150], with a minimum number of two sources (see Methods). Only the miRNA and predicted target mRNA pairs for which there was a statistically significant negative Spearman correlation across at least 5 cancer types were considered, and in doing so this identified the miRNA-target pairs showing statistically significant recurrent negative correlation in expression across cancer types by the rank product test (Figure 3.4a). As depicted by the process in Figure 3.4b-c, analysis of these statistically significant miRNA-target pairs revealed a strong enrichment for tumour suppressor

genes (as defined by the COSMIC database list of 141 TSG), as might be expected for miRNA associated with oncogenic processes (p = 0.0006, two-sided Fisher's exact test). Likewise, oncogenes were not found to be enriched within this list of genes; 22 were found to overlap (p = 0.11, two-sided Fisher's exact test).

A different picture emerged after this analysis was repeated for oncogenes, and for the miRNAs found to be statistically significantly negatively associated with one or more hallmark signature across cancer types. 1283 statistically significantly anti-correlated miRNA-target pairs were identified with the rank product test across cancer types for recurrently negative Spearman correlation, for the downregulated hallmark-associated miRNAs. Among these 1238 genes, oncogenes were enriched, with 42 oncogenes present in this list of the 231 COSMIC oncogenes, which was statistically significant by two sided Fisher's exact test ($p < 10^{-5}$). Likewise, this list was not enriched statistically significantly for tumour suppressor genes, as just 14 were found to overlap (p = 0.08, two sided Fisher's exact test). Next, all predicted miRNA-oncogene interactions among the 231 COSMIC oncogenes were analysed, and among these, only 2 showed statistically significant anticorrelation across tumour types with their predicted target miRNA (*ESR1* and *ABL2*). The intersection of these lists of 2 COSMIC oncogenes and the 1283 miRNA-oncogene pairs associated with gene signatures was taken, and identified only *ESR1* (interacting with miR-18a-5p and miR-130b-3p) in common ($p = 1.2 \cdot 10^{-5}$, two-sided Fisher's exact test). This suggested that *ESR1*, estrogen receptor alpha, may be involved with the hallmarks of cancer, or may be a feature of the data analysed, as the number of samples considered is skewed towards breast tumours, where *ESR1* is known to play a key biologic role [151, 152].

3.3.5 A core set of tumour suppressor genes shows statistically significant association with the hallmark gene signatures across cancer types

Next, an analysis of the miRNA-associated tumour suppressor genes that showed statistical significance in downregulation, in the context of all other tumour suppressor genes, was carried out, as the initial selection of miRNA, and therefore their predicted targets, was biased by the gene signatures chosen. That is, the above analysis was repeated with every predicted miRNA-TSG pair, for each TSG independently, considering again the statistically significant associations across at least 5 cancer types, and then these statistically significant associations were collated across cancer types with a rank product test, as summarised in Figure 3.4d. This second portion of the

analysis revealed the miRNA statistically significantly anticorrelated with each TSG across cancer types, thereby acting to mitigate the bias accumulated by comparing regulation of multiple TSG in the same analysis. Then, considering the miRNA-TSG pairs found to be statistically significantly anticorrelated across cancer types in both analyses from Figures 3.4c and d, a set of 22 miRNA-TSG pairs was identified in common, which was comprised of 8 TSG (*FAT4*, *TGFBR2*, *ARHGEF12*, *DNMT3A*, *CDK12*, *ACVR2A*, *SFRP4*, and *PTEN*) and 17 miRNA, as shown in Figure 3.4e. In addition, the miRNA found to be associated to each of these TSG were, in many tumour types, expressed at statistically significantly higher levels in wildtype cases for the associated TSG (Figure B.10, Appendix B4).

3.3.6 The statistically significant associations of signature-associated miRNA differs between tumour and adjacent normal samples

The presented analysis highlighted the statistically significant associations of miRNA to the gene signature scores across cancer types. An analogous analysis in normal tissues was carried out to assess the differences that may be present for miRNA statistically significantly associated with gene signature scores within adjacent normal tissues. Tissue types with at least 20 adjacent normal samples with both miRNA and mRNA expression data, from the TCGA dataset were obtained. This yielded data from 6 tissue types: BRCA, UCEC, HNSC, KIRC, LUAD, and BLCA, and for each of these, a linear model was fit, with response variable as the gene signature score and using miRNA expression as the set of predictors. The model was fit using the combined univariate-multivariate penalised linear regression approach as used for the fitting of miRNA to tumour gene signature scores. As in the previous analysis, coefficients for the miRNA were aggregated across tissue types, and those showing consistent high or low ranking, using the rank product test, with statistical significance defined as Bonferroni p value less than 0.05.

In this way, a set of miRNA statistically significantly associated both positively and negatively with the hallmarks gene signatures was identified in normal tissues. The overlap of these miRNA with the miRNA identified through the analysis of tumour samples was statistically significant for each signature considered. On average, an overlap of 54% of miRNA was observed for miRNA positively and negatively associated with signatures, which was highly statistically significant ($p < 10^{-19}$ in all cases, by two-sided Fisher's exact test).

Although fewer normal samples were considered in analysis (minimum of 20 samples required versus 200 samples minimum for tumour tissues), the validity of gene signature application to these datasets was assured using *sigQC*. This analysis, depicted in Figures B.1- B.9, showed that normal datasets were comparable in quality for the application of these gene signatures to tumour datasets.

Examining the predicted targets of these positively signature-associated miRNA from normal tissues, 233 recurrently negatively correlated miRNA-target pairs were identified, of which two contained miRNA-TSG pairs (*CEBPA* and *NCOA4*). However, this overlap of the 142 unique genes among the 233 miRNA-target pairs with the 141 COSMIC tumour suppressor genes did not show statistical significance ($p = 0.26$ by two-sided Fisher's exact test). That is, while very similar miRNA were found to associate statistically significantly with each gene signature, regardless of tumour or normal dataset, the predicted targets of the miRNA with recurrent negative correlations across cancer types showed less overlap with tumour suppressor genes among adjacent normal samples as compared with tumour samples. This differential statistically significant association may be indicative of the underlying biological difference between these samples, but the degree to which these correlations show effect in reality must be experimentally determined.

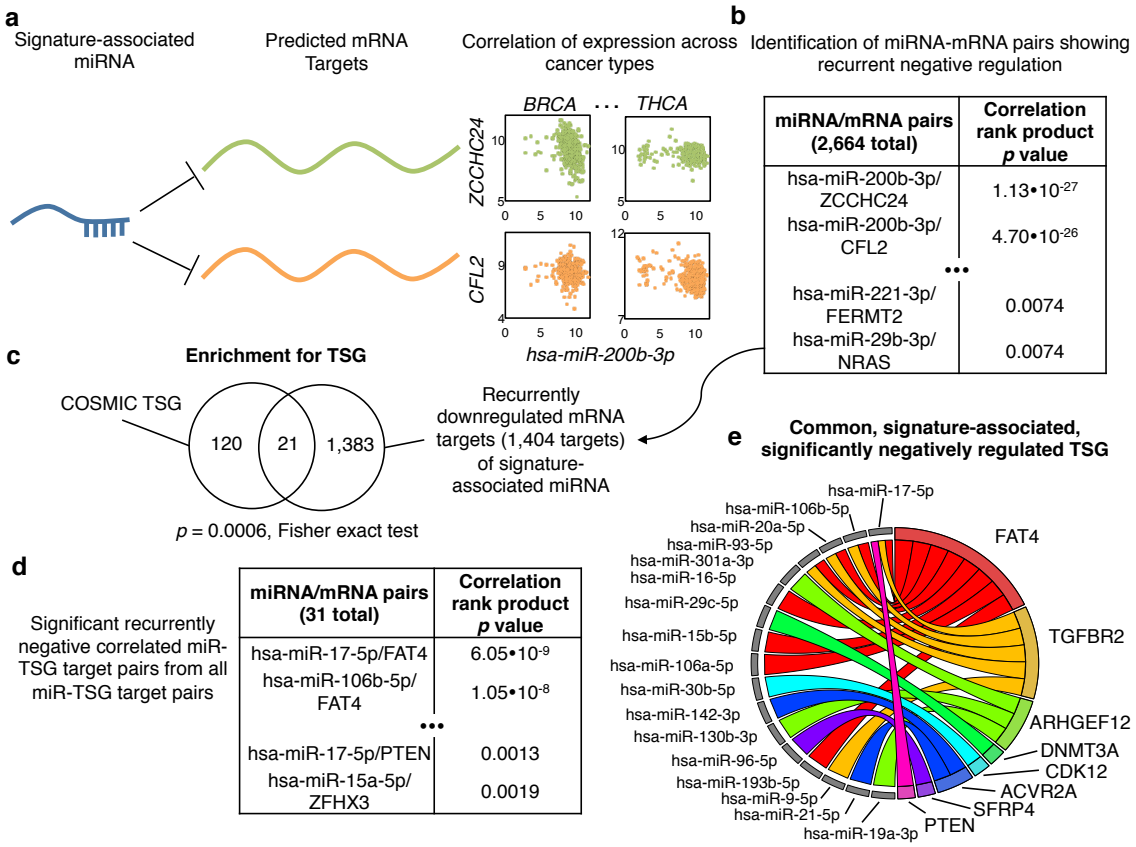

Figure 3.4: **Approach used for interpreting miRNA-target interactions.** (**a**) First, miRNA-target pairs for each positively associated hallmark-associated miRNA were identified, and the correlation between these was determined. (**b**) Next, the correlations across cancer types were aggregated, and those identified as consistently negative-ranking were identified with the rank product statistic. (**c**) Among this list of miRNA-mRNA target pairs, there was statistically significant enrichment for tumour suppressor genes, as identified by the two-sided Fisher exact test. (**d**) The same procedure as described in (**a**) and (**b**) was repeated for all miRNA and all predicted target TSG pairs. (**e**) From the lists identified in (**b**) and (**d**), those miRNA-TSG pairs in common were identified, and their interactions were displayed on a circos plot, showing the repressive actions of each miRNA on its predicted target TSG.

3.3.7 The statistically significant associations of signature-associated miRNA differs between breast cancer subtypes

Differential miRNA statistically significant association to targets was also detected among different breast cancer subtypes. The basal breast cancer subtype and luminal B breast cancer subtypes, as defined for the TCGA breast cancer dataset were considered. In a method analogous to that presented for previous analysis, a linear model describing the relationship between miRNA expression and gene signature score for the 24 hallmarks signatures was fit, using penalised linear regression and ten fold cross-validation. This was done independently for the two breast cancer subtypes considered. Once the linear models were fit and coefficients for each miRNA were obtained, for each miRNA, the difference in coefficient between the two breast cancer subtypes, for each signature, was taken. These differences were then compared across the 24 signatures, and those consistently higher or lower than expected due to chance were identified using the rank product statistic. The miRNA identified through this analysis therefore represented those that were differentially statistically significantly associated with the gene signature scores representative of the hallmarks of cancer, between breast cancer subtypes. As a visualisation of this context-dependence, the ten most associated miRNA to basal breast cancers exclusively and the ten most associated to luminal B breast cancers exclusively were plotted, as depicted in Figure 3.5. Further interrogating these context-dependencies, with a greater number of samples, and therefore statistical power, may help to uncover unique biology between tumour subtypes, as these observations may be due to statistical noise or lack of samples, and require experimental validation for confirmation.

3.3.8 Analysis of modes of regulation reveals copy number and mutational status are key determinants of TSG expression

With a set of TSG potentially regulated by miRNA in relation to phenotype identified, the determinants of their expression were characterised next. Methylation status, copy number, miRNA expression, and mutational status (see Methods), were used as predictors for the linear model depicted in Figure 3.6a. Notably, when considering the miRNA in this model, all reported miRNA were considered to potentially add evidence for novel miRNA-target associations. This model was then fit using penalised linear regression over the various cancer types, and then coefficients were aggregated across cancer types using the rank product statistic. This identified statistically

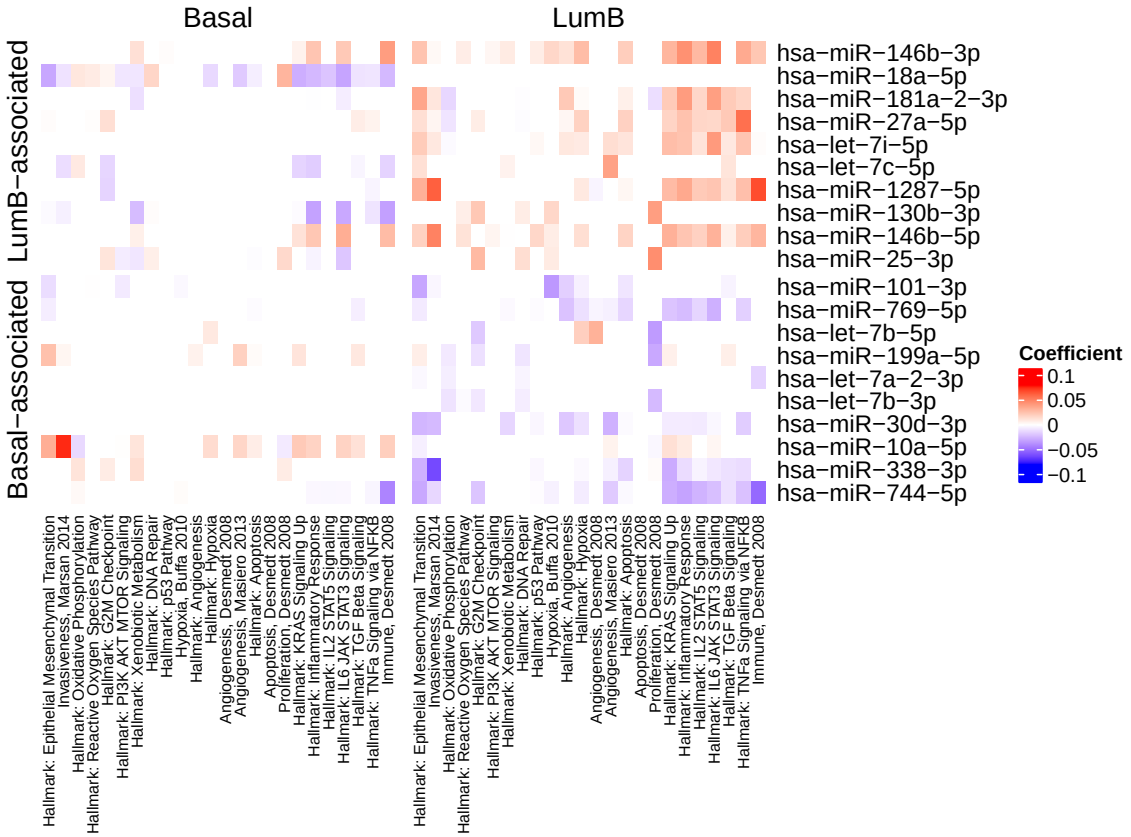

Figure 3.5: **Differential statistically significant associations of miRNA in breast cancer subtypes.** Heatmap depicting statistically significantly different miRNA associated to hallmarks gene signatures between basal and luminal B breast cancer subtypes.

significant recurrent positive and negative predictors of expression across cancer types individually, for each of the 8 tumour suppressor genes identified in Figure 3.4e. This analysis yielded both expected results, such as the important positive predictive role of copy number for each of the tumour suppressor genes, as seen in the left panel of Figure 3.6b, and unexpected associations, such as the statistically significant positive association of many miRNA and methylation probes with TSG expression in some cases. Positively-associated miRNA likely arose in this analysis as a result of the inclusion of all miRNA expressed in each cancer type, as opposed to those only predicted to target the TSG, so that novel associations could be hypothesised. The positively associated miRNA appeared to be co-expressed for a variety of reasons, such as competitive endogenous mRNA interactions, repression of repressors of the TSG, or presence on a nearby genomic locus, subject to the same enhancer or promoter. It is important to note that this approach in using penalised linear regression worked to

minimise the effects of miRNA present on a nearby genomic locus, as copy number was included as a covariate in the linear model. That is, because copy number and positive miRNA associations are correlated with each other, the penalised regression worked to ensure that the effect of similar genomic location was represented by the statistically significant association with copy number, and not a co-localised miRNA, but this effect could not be completely annulled.

Likewise, the identified modes of negative regulation gave expected results, with nonsense mutations and frameshift deletions consistently negatively statistically significantly associated with TSG mRNA expression. Further, because this analysis was done with all miRNA, and not just those predicted to have a given TSG target, these results allow for the hypothesis of novel miRNA-TSG interactions, represented by the recurrently negatively correlated miRNA with each TSG, but confirmatory experiments are required before any such conclusions can be drawn. The complete rank product tables and all autocorrelation matrices can be found in Appendix B5.

3.3.9 *PTEN*, *FAT4*, and *CDK12* expression show strong statistically significant association with either miRNA, promoter methylation, or mutation across tumour types

Once the potential modes of regulation and their relative importance were estimated (Figure 3.6), the relative occurrence of each of these modes of regulation was determined. The negative regulators which co-occurred were identified, and conversely those which were exclusive repressors were identified (Figure 3.7a). Preliminary analysis of autocorrelation heatmaps (e.g. Figure 3.7a) revealed that in some cases, the statistically significant negative association with miRNA was nearly exclusive from the statistically significant negative association with methylation probes. A full series of heatmaps for all cancer types considered and all tumour suppressor genes with their associated candidate negative regulators identified is presented in Appendix B6, Figures B.12- B.19. These results were consistent with the possibility of regulation of TSG expression by either miRNA or methylation, in addition to deletion or mutation. To characterise this, a bootstrap resampling based approach was devised, to determine the statistical significance of the difference in co-correlation between the miRNA and the methylation probes themselves, and then with each other. For each cancer type, the significance value of this proportion was calculated (Figure 3.7b). This suggested that for each of the TSG considered, there were tumour types for which the statistically significant associations were consistently exclusive. Further, it also

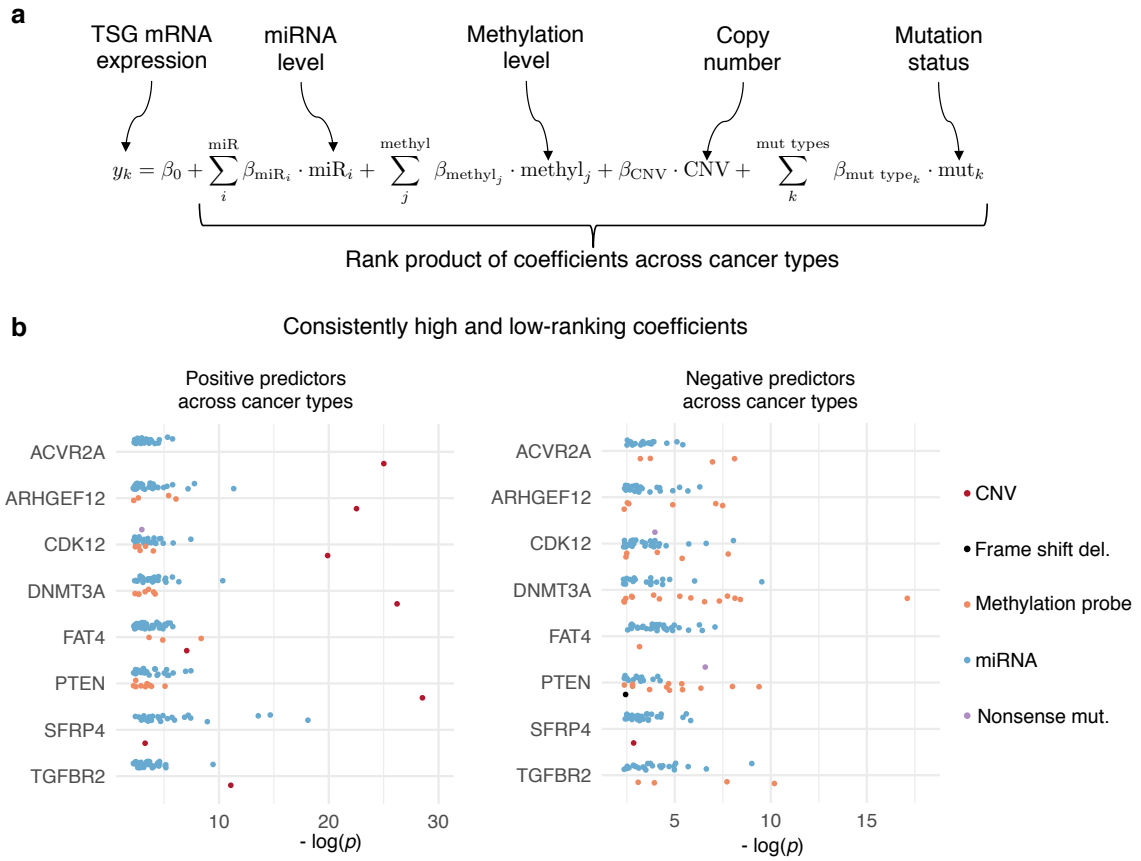

Figure 3.6: **Approach used in determining the regulation of each TSG identified as potentially statistically significantly miRNA-regulated.** (**a**) The linear model used whilst determining predictors of TSG mRNA expression. (**b**) Model coefficients were aggregated across cancer types with the rank product statistic, and those identified as statistically significant positive and negative predictors are depicted alongside the -log of their rank product p-value.

arose that across multiple tumour types, the genes *PTEN*, *FAT4*, and *CDK12* consistently tended towards exclusivity in their statistically significant associations with possible negative regulators, lending support for the importance of possible miRNA-based regulation of these tumour suppressor genes. More specifically, for *PTEN*, the difference in proportions of positively co-correlated miRNA and correlated miRNA and methylation probes, was above what would be expected due to chance, more than 80% of the time as compared to bootstrapped runs, across nearly every cancer type considered, and the same was observed for *CDK12* and *FAT4*, over 90% of the time, in most cancer types, as shown in Figure 3.7b.

The calculations for the analysis of TSG regulation and analysis for the exclusivity of gene regulation were repeated for an independent dataset comprising matched

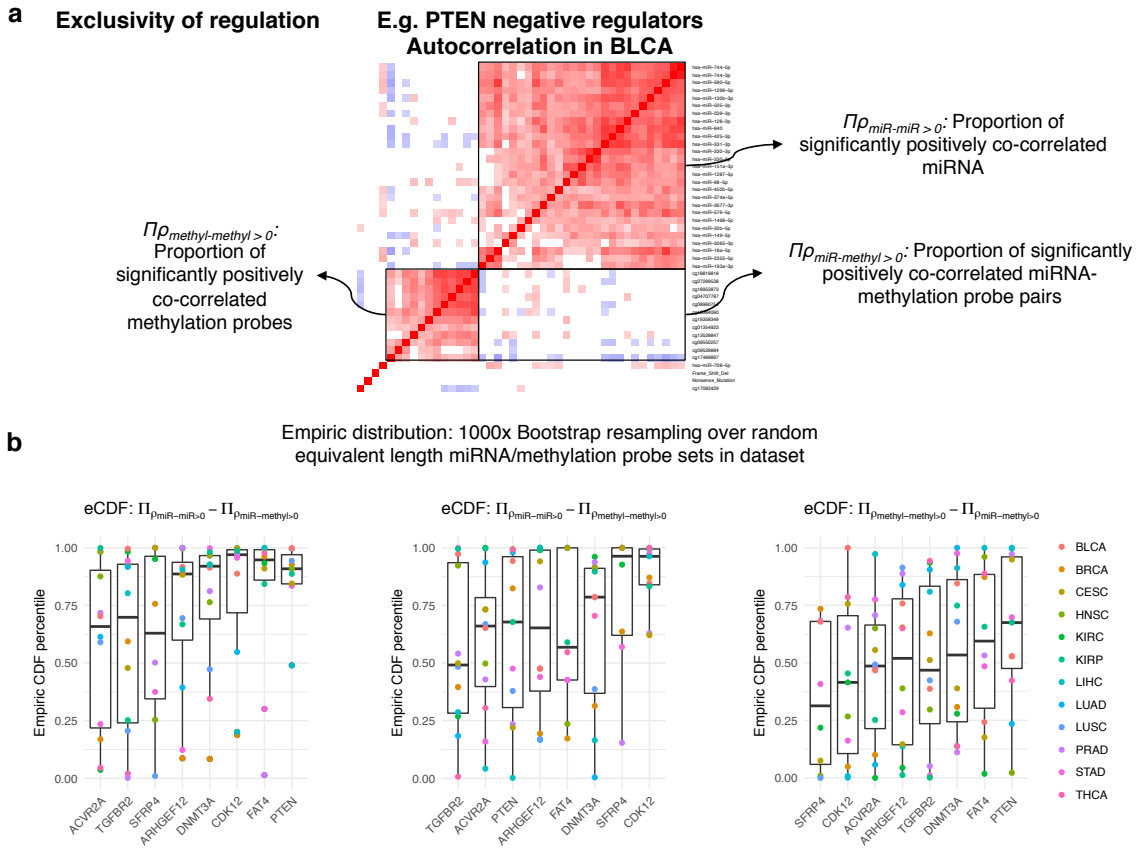

Figure 3.7: **Summary of exclusivity of TSG regulation by miRNA, showing trends towards exclusivity of statistically significant association of negative regulators across cancer types for *PTEN*, *FAT4*, and *CDK12*.** (**a**) Depiction of the autocorrelation heatmap for the expression of the various negative regulators of the tumour suppressor gene, and the variables considered and their meaning, as depicted. (**b**) Plots depicting the spread of the percentiles on the empiric CDF for the distributions for the pairwise differences of the variables identified in (**a**) through a bootstrapping-based analysis, as described in the Methods section.

mRNA, miRNA, copy number variant (CNV), mutation, and methylation data for 93 patients with ovarian cancer, from the OV-AU project from the ICGC data portal [125]. Results of this analysis are highlighted in Figure 3.8 below.

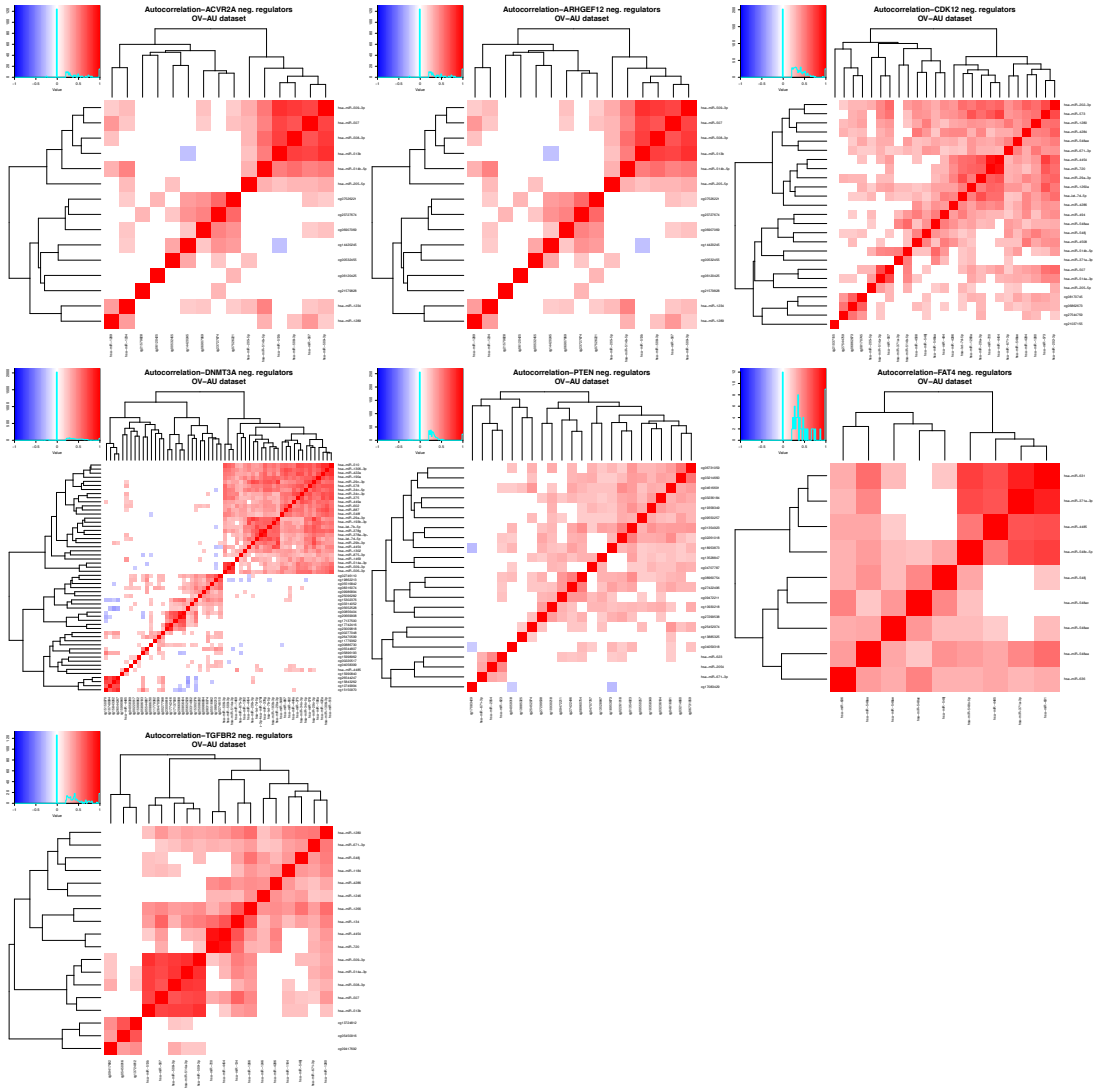

Figure 3.8: **Evidence for independent regulation can be seen for the same TSG and negative regulators in an independent ovarian cancer dataset.** Autocorrelation heatmap for the expression of the identified negative regulators of each of the 8 statistically significant TSG identified in this study in an independent ovarian cancer dataset. *DNMT3A* ($\Pi_{\rho_{\text{miR-miR}}} - \Pi_{\rho_{\text{miR-meth}}} = 0.11$, $\Pi_{\rho_{\text{miR-miR}}} - \Pi_{\rho_{\text{meth-meth}}} = 0.01$, and $\Pi_{\rho_{\text{meth-meth}}} - \Pi_{\rho_{\text{miR-meth}}} = 0.99$) and *PTEN* ($\Pi_{\rho_{\text{miR-miR}}} - \Pi_{\rho_{\text{miR-meth}}} = 0.18$, $\Pi_{\rho_{\text{miR-miR}}} - \Pi_{\rho_{\text{meth-meth}}} = 0.05$, and $\Pi_{\rho_{\text{meth-meth}}} - \Pi_{\rho_{\text{miR-meth}}} = 1$) tend towards exclusivity in this dataset.

Next, the statistically significantly negatively associated miRNA and methylation probes, along with mutation status, were used to define subgroups of samples, for which decreased TSG expression was shown, in the subgroups with high expression of these miRNA or high methylation of these probes, as depicted in Figures B.20- B.29 in Appendix B7. Further, these miRNA high and highly methylated samples had transcriptomes altered in a similar manner as in TSG mutated cases, as was established via an analysis of differentially expressed genes in both cases, with statistically significantly positive Spearman correlation for fold change across all genes in every case considered, as shown in Figures B.30- B.38 in Appendix B7.

3.3.10 *ARHGEF12*, *SFRP4*, *TGFBR2*, and statistically significantly associated miRNA show strong association with breast cancer molecular subtype

Next, statistically significant associations between TSG and tumour molecular subtypes were identified. An analysis of the eight identified tumour suppressor genes consistently negatively associated with miRNA across cancer types showed that in many cases, their mRNA levels were associated with breast cancer molecular subtype. In particular, the basal subtype showed the lowest median expression of *ARHGEF12*, *SFRP4*, and *TGFBR2*, as compared to normal tissue, luminal A, B, Her2 amplified, or normal subtypes of breast cancer as shown in Figure 3.9 below, and this association was statistically significant even when cases were restricted to wildtype expression of *ARHGEF12*, *SFRP4*, and *TGFBR2*. At the level of the statistically significantly associated miRNA identified as potential negative regulators of these TSG, it was shown that the median expression of these miRNA was also statistically significantly associated with breast cancer molecular subtype ($p < 0.05$, multiple associations, identified by Wilcoxon and Kruskal-Wallis testing), and this relationship was reversed for TSG mRNA expression by subtype, as expected. It was also shown that these statistically significant associations were preserved when samples with non-silent mutations in the TSG were removed. For further validation, it was shown that these TSG and miRNA statistically significant associations to breast cancer subtype were reproducible in the independent Metabric dataset ($N = 1293$) [124].

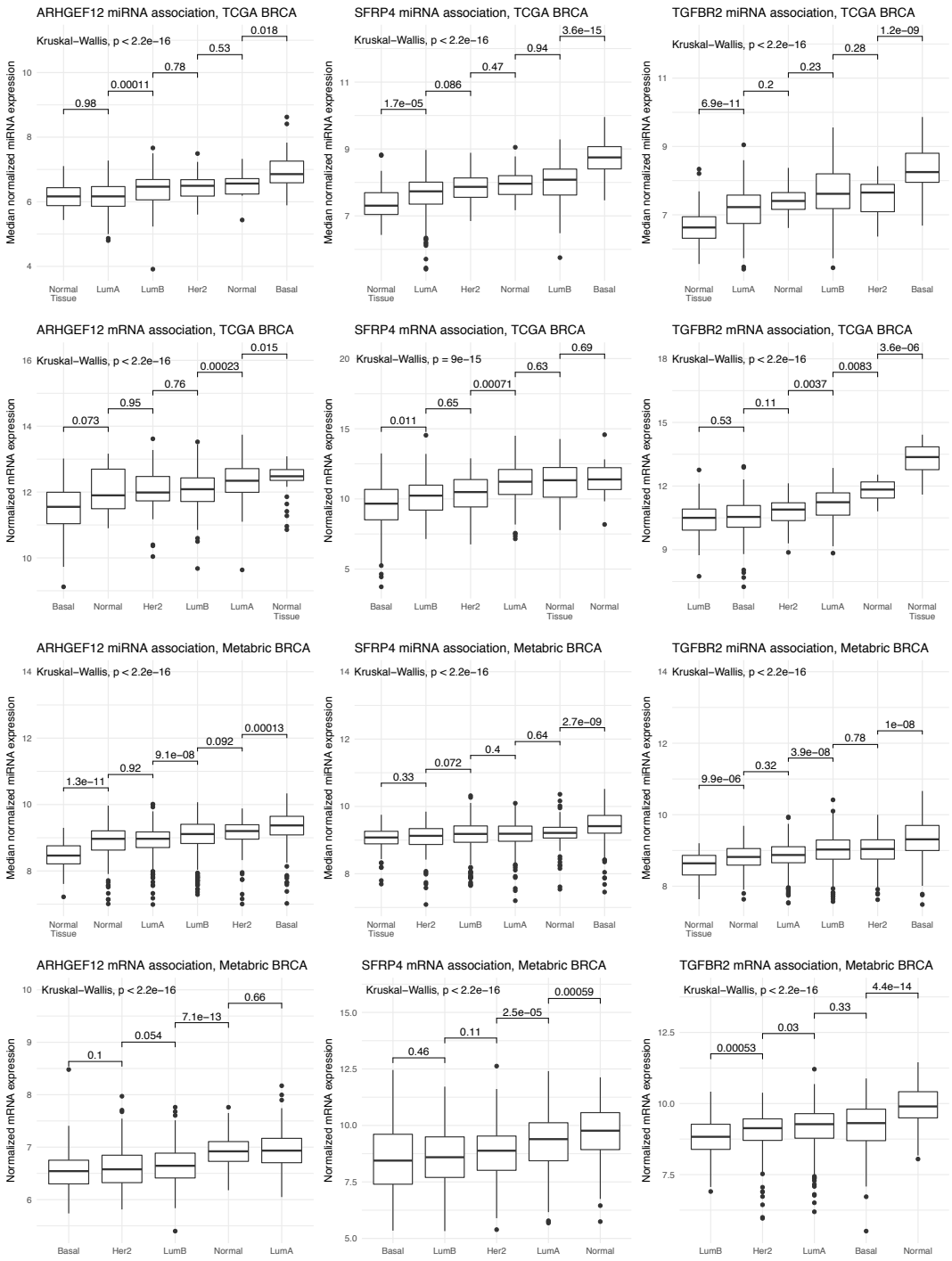

Figure 3.9: ***ARHGEF12*, *SFRP4*, and *TGFBR2* expression associates statistically significantly with breast cancer subtypes in TCGA and Metabric cohorts.** Panels displaying the median normalised miRNA expression and normalised mRNA expression level associations of tumour suppressor genes and breast cancer molecular subtypes. Columns, left to right, represent plots for *ARHGEF12*, *SFRP4*, and *TGFBR2* respectively. The top two rows show median miRNA and mRNA expression for the TCGA breast cancer dataset, respectively. The bottom two rows show median miRNA and mRNA expression for the Metabric dataset, respectively.

3.4 Discussion

3.4.1 A first large-scale association of miRNA to hallmarks of cancer

In this chapter a rigorous statistical analysis of the statistically significant associations between miRNA and mRNA gene expression signatures was carried out. Gene signatures represent transcriptomic association utilised in two key ways, which strengthened the approach. First, the use of gene signatures to understand the relationship of non-coding RNA and phenotype relies on the phenotypic associations intrinsic to quality-controlled gene signatures. Second, because miRNA can only repress mRNA that are present in sufficient quantity in a cell, when inferring function, it is vital to group transcriptomic profiles by miRNA targeted gene expression. This allows for an understanding of the possible miRNA-associated gene regulation important to the phenotype one wishes to uncover. Thus, this analysis represents a novel approach to understanding the complexity of miRNA association to phenotypes, which is particularly relevant in the context of cancer.

This chapter began with ensuring applicability of the gene signatures, and then for each signature, the miRNA both statistically significantly up- and down-regulated in association with the signature score were identified. From this, the network shown in Figure 3.1 was obtained, which describes in a detailed fashion, across tumour types, the many statistically significant associations of individual miRNA with each phenotype. Reproducibility of this network was also shown in an independent dataset, by considering the overlap with the network identified using the Metabric dataset and the same gene signatures.

Moreover, repeating this analysis, grouping the miRNA statistically significantly upregulated and downregulated by miRNA family yielded the surprising result that many miRNA families were found to statistically significantly associate in opposing directions with the hallmarks of cancer; including 4 of the top 5 most common miRNA families involving the miRNA found to be associated by this analysis (miR-25 family, miR-17 family, miR-15abc family, and let-7 family). By virtue of the high similarity between their seed regions, often resulting from a common evolutionary ancestry, miRNA families are thought to consist of miRNA with similar biological function redundantly targeting the same mRNA [153].

This observation challenges the prevailing hypothesis of miRNA families associating in a coordinated fashion across multiple phenotypic states, and suggests possible statistically significant differential association of individual miRNA depending on cell

type for which it is expressed, regardless of grouping by family [120, 148, 149, 154, 155]. However, as noted above, it is crucial to temper this observation by noting that these statistically significant associations in different directions may have arisen due to variation or noise, and require experimental confirmation. Future studies with larger cohorts will be required to confirm these results.

Related to the discussion of statistical power, is the discussion of the validity of the statistically significant miRNA-target interactions inferred from various target prediction algorithms. In this study, miRNA-target gene interactions were predicted using the miRNAtap database in R, version 1.14.0, as described in the Methods section. While this analysis has not been repeated further to include greater stringency in target selection (e.g. requiring a miRNA-target interaction predicted by more than two independent sources to be considered), the analysis methodology itself was designed to obtain high-confidence targets. Initially, many possible miRNA/predicted mRNA targets were retained, with a balance of potential false-positives, requiring a minimum of two sources predicting the interaction. In this way, a reasonable number of interactions were included, while not being restrictive to those which were predicted by the commonalities of each algorithm. With this more comprehensive list, it was tested directly which of these miRNA/predicted target pairs themselves showed potential repressive, statistically significant, association using correlations in the data, and then this list was further refined using the rank product statistic. Thus, for the analysis presented, there is initially a reasonably wide and comprehensive list of potential targets, which are refined using the power afforded by the large datasets considered.

3.4.2 miRNA associate in coordinated networks to potentially achieve functional effect

As might be expected, given the complexity of the action of non-coding RNA, it was shown that for a given phenotype, single miRNA-predicted target associations did not account for gene signature-associated phenotype; rather it was the case that changes arising from a network of miRNAs, possibly interacting with many targets in a coordinated manner, serve to modify the transcriptome to achieve the phenotypes associated with cancer. That is, because the targets of a given miRNA are predicted to be variable in their function, and are not all present in every sample at repressible concentrations (often necessitating the use of an expression level filter), the same miRNA may be statistically significantly associated with opposing phenotypic effects

in different contexts, as reported by Denzler et al. in [121, 156] for competing endogenous mRNA (ceRNA). ceRNA have also been identified in a high-throughput fashion by Chiu et al. in [122] and Xu et al. in [123] and an approach to their identification, describing the necessary experimental and statistical prerequisites has been reported by Smillie et al. in [157], with recent work identifying such networks involving *PTEN*, for instance [158]. These observations are also supported in the case of triple negative breast cancers as compared with all ER-negative tumours, in an analysis by deRinaldis et al. [159]. In this study, the authors considered 181 breast tumour clinical samples, and identified 7 prognostic miRNA in triple-negative breast tumours, and a further 7 prognostic miRNA in all ER-negative tumours [159]. The miRNA found to associate with poorer prognosis were associated with transcriptional programs linked to the hallmarks of cancer, as was identified by the results presented above [159]. Indeed, it was shown that the expression of these miRNA was linked to various oncogenic programs, such as cell-adhesion, motility mechanisms related to tissue invasion, and growth factor-mediated signalling pathways [159]. It was shown in this section that the statistically significant association of miRNA to predicted targets may be dependent on the tissue type, the underlying genomic aberrations present, and molecular subtype of breast cancer they were expressed in, though this requires further experimental validation to validate. Moreover, if tissue type does impact miRNA behaviour, sample purity is a crucial cofactor to consider when studying miRNA and predicted target statistically significant association. Further study into deconvolution methodologies enabling more accurate quantification of miRNA abundance from purely tumour samples will likely elucidate miRNA-predicted target interactions more clearly. Such deconvolution-based methods would need to ensure that both the expression of miRNA and mRNA are corrected for purity before testing for any such correlations, using methods such DeconRNASeq [160] or Cibersort [161]. Alternatives to these methods may involve removing samples with low estimates of tumour purity, or removing the miRNA with strong correlations to tumour purity from analysis.

3.4.3 Implications for miRNA-based therapeutics

As miRNA are increasingly thought of as potential therapeutic agents, if miRNA are to have effective therapeutic function, a single miRNA may be an ineffective strategy. Rather, a cocktail of miRNA may be needed to alter the cancer cell transcriptome. This has been shown *in vitro*, with comparisons of multiple miRNA versus a single miRNA targeted highlighting that when multiple miRNA with co-ordinated function

are modified, greater phenotypic change is seen, as reported in multiple publications [162, 163, 164, 165, 166]. Potentially, for miRNA therapeutics to achieve function, these may have to be based on a number of miRNA, given to a highly selected group of patients with transcriptomes deemed to be potentially responsive to this network perturbation. Further, by using more than a single miRNA as a therapeutic agent, the off-target effects that have limited development in this field may be mitigated, by buffering for this with other miRNA in off-target tissues [167, 168, 169, 170]. In fact, a recent Phase I trial of delivering miR-34a to patients with solid tumours identified acceptable safety overall, but many patients in this trial suffered from off-target immune-related toxicities, such as fever, lymphopenia, and neutropenia, necessitating pre-treatment with dexamethasone [171]. However, this field is in its infancy, and will absolutely require further study before even considering clinical translatability, and these ideas are offered as potential avenues to be explored.

A potential implication raised by the findings in this section is that there are tumour suppressor genes for which a mutation is not requisite for inactivation, but rather, for which inactivation is achieved through miRNA-mediated repression or methylation-mediated repression alone. For the TSG identified as miRNA associated statistically significantly, it was also shown that TSG mutations occurred independently of *MYC* amplification, which itself was recently identified as an independent regulator of miRNAs. In addition, it was shown that such *MYC* amplification status was indeed associated with miRNA expression for the miRNA found to be negatively associated with each of the TSG in a majority of cases (Figure B.39, Appendix B8). Further, it was shown that in particular tumours, for *PTEN*, *CDK12*, and *FAT4*, this putative miRNA or methylation-based suppression appears to occur statistically significantly independently of other gene regulatory factors, such as mutations and copy number changes.

Lastly, the ability of the presented approach in this section to potentially capture tumour biology was highlighted through the identification of tumour suppressor genes showing recurrent statistically significant anticorrelation with cognate miRNA across tumour types, which were shown to have statistically significant association in expression to breast cancer molecular subtype. Specifically, this analysis pointed towards the role of decreased mRNA levels of *ARHGEF12*, *SFRP4*, and *TGFBR2* as associated with basal breast cancer [172, 173]. Having identified potential negative regulators of these TSG, it was next shown that the corresponding miRNA alone associated with breast cancer subtype, elevated in the basal subtype, possibly representing novel biological association, though this requires experimental validation.

The presented methodology may be used in future work encompassing both more specific signatures, as well as larger, more expansive datasets to derive increased confidence in particular statistically significant associations uncovered.

3.5 Summary and Conclusions

In this chapter a methodology was developed by which miRNA that statistically significantly associate with mRNA gene expression signatures, across large datasets could be identified. This methodology was used to elucidate the miRNA associated with 24 quality-controlled gene signatures representative of the hallmarks of cancer among 15 epithelial cancer types from the TCGA dataset. Specifically, evidence was generated for the anticorrelation of *PTEN*, *FAT4*, and *CDK12* to miRNA in samples, showing statistical evidence of exclusivity to negative correlation with other modes of regulation, such as mutation, deletion, and promoter methylation. It was shown that the miRNA themselves showing repressive ability for three TSG; namely, *ARHGEF12*, *SFRP4*, and *TGFBR2*, were able to associate independently with the molecular subtypes of breast cancer. In conclusion, through the work presented within this chapter, it has been shown how one may capture novel miRNA statistically significant associations with mRNA and phenotype, across multiple datasets, generating hypotheses that lend themselves towards experimental validation.

Chapter 4

Hypoxic breast cancers may display a *DICER1*-independent, *AGO2*-dependent miRNA biogenesis pathway

Abstract

In this chapter, I examine the expression, mutational, and copy number changes in the genes involved in miRNA biogenesis, as they vary with the hypoxia gene signature score across 15 epithelial cancer types. These findings point towards global changes in miRNA levels and maturation in hypoxia, occurring as *DICER1*, *TNRC6*, *DDX5*, and *DDX17* are down-regulated and *AGO2* and *PABPC1* are upregulated, particularly among hypoxic breast tumour clinical samples. These global changes may arise because of a shift in the miRNA biogenesis pathway, leading to differential processing and maturation of specific miRNA. Moreover, I show that the miRNA that appear to be selectively matured in the hypoxic microenvironment are those that repress tumour suppressor genes, which may facilitate adaptation to hypoxia. The results within this chapter present evidence suggesting novel hypotheses related to the fundamental question of how cancer cells modify their phenotype in order to adapt to the harsh hypoxic microenvironment.

4.1 Introduction

4.1.1 The canonical miRNA biogenesis pathway

As shown in previous chapters, miRNA possess the key property of inflicting potentially large-scale context-dependent changes upon the transcriptome. Their function in modifying the state of the transcriptome is one that has been shown to potentially associate with carcinogenesis, through putative repressive actions on tumour suppressor genes. As a result, in the search for an understanding of their role in cancers, I next examine the enzymes and cofactors regulating the production of these biomolecules. The biogenesis of miRNA is a tightly regulated process within the cell, and much work has been done to characterise a canonical biogenesis pathway and define the biomachinery genes involved in this pathway.

In brief, the canonical pathway for miRNA biogenesis begins with transcription of the gene encoding the miRNA from the genome by RNA polymerase II [174]. This pri-miRNA transcript, is then cleaved into a 70-nucleotide stem-loop sequence known as a pre-miRNA, within the nucleus by the microprocessor complex, comprised of the enzymes *DROSHA* and *DGCR8* in humans [174]. This stem-loop structure is then exported from the nucleus into the cytoplasm with the help of Exportin-5 (*XPO5*) [174]. Once in the cytoplasm, the stem-loop structure is cleaved again into a duplex of two 22nt single-stranded mature miRNA by the enzyme *DICER1* into miRNA with a 3' and 5' overhang, termed -3p or -5p, respectively [174]. This duplex then unwinds whilst the RISC protein machinery is assembled at a cellular miRNA processing site, known as the P-body [174]. Within the RISC, one of the mature strands binds to the seed region of the target mRNA, and the other degrades, and then in conjunction with *AGO2*, translation is repressed or the mRNA is degraded [174].

4.1.2 Alternative miRNA biogenesis pathways

In addition to the canonical pathway, evidence is now emerging for alternative biogenesis pathways in complex organisms. For instance, it has been shown that through an interaction with *AGO2*, the elongation factor eIF1A functions to promote *DICER*-independent biogenesis of miRNA [175]. Yi et al. show that in zebrafish, through this *DICER*-independent biogenesis, miR-451 is preferentially produced and processed, and acts to rescue the organism from a state of knocked down or mutant eIF1A protein [175]. That is, through preferential production of a specific miRNA interacting with eIF1A, Yi et al. showed that an alternative miRNA biogenesis pathway was able to rescue the organism from this perturbed state [175].

At the level of human cell lines, recent work by Kim et al. studied the potential for such alternative pathways, using small interfering RNA (siRNA)-based knockout experiments [176]. In this work, Kim et al. create cell lines with knocked out *DROSHA*, *DICER1*, and *XPO5*, and characterise the non-coding transcriptome in each case [176]. They observed that *DROSHA* knockouts have the most severe effect on the reduction of the miRNA expression levels, and results in the loss of nearly all miRNA expression, whereas knockout of XPO5 only resulted in a modest reduction in miRNA production, suggesting it is not critical for the biogenesis of miRNA [176]. In the case of *DICER1* knockout, Kim et al. observed that *AGO2* is able to compensate for the slicing functionality of *DICER1*, and as a result process miRNA into mature form in a *DICER1*-independent manner [176]. In doing so, Kim et al. showed that the slicing activity of *AGO2* preferentially produces -5p miRNA and leaves a 3' overhang on these processed miRNA [176]. Lastly, Kim et al. showed that the effects of the compensatory production of miRNA through *AGO2*-mediated slicing, *DICER1*-independent biogenesis extend only to specific miRNA, but do not provide further insight into the functional roles of these miRNA, or situations in which production of these miRNA might be selected for by evolutionary forces [176].

4.1.3 Regulation of miRNA biogenesis involves multiple levels of control

While a cell may respond to changes in its environment relying on alternative miRNA biogenesis pathways, it may also respond through regulation of the canonical miRNA biogenesis pathway, as summarised in a recent review by Ha and Kim published in Nature Reviews Molecular Cell Biology [177]. At each level of the multistep process comprising miRNA biogenesis, there is the opportunity for other proteins to interact with the core components, regulating their behaviour [177]. These companion proteins themselves are subject to control by methylation, copy number changes, post-transcriptional, and post-translational level control [177]. This fine regulation of the biomachinery genes has the potential to closely modulate both the global level of miRNA across the transcriptome, and the maturation of specific miRNAs [177]. For example, DROSHA and DGCR8 constitute the core members of the microprocessor complex, responsible for the cleavage of pri-miRNA into pre-miRNA within the nucleus [178]. For the microprocessor, it has been shown that the binding of DDX17 protein to a specific motif of the pri-miRNA nucleotide sequence enhances the activity of the complex, adding a further regulatory dimension to pri-miRNA processing [178]. There are many such partner genes, and in this chapter I consider a

panel of the core biogenesis genes, and companion genes which have been validated in altering the behaviour of the miRNA biogenesis pathway, summarised in Table 4.1.

4.1.4 Known changes in miRNA biogenesis in hypoxia

Recent work has shown that hypoxic tumours are associated with a global decrease in miRNA biogenesis, and studies of the biogenesis pathway reveal hypoxia-associated reductions in the expression levels of *DROSHA* and *DICER1* in cell lines [179, 180]. Rupaimoole et al. showed that *DROSHA* is decreased in hypoxia by interactions with the ETS1/ELK1 transcription factors, and van den Buecken et al. showed that in hypoxia, *DICER1* is epigenetically repressed by repression of the demethylases KDM6A/B for its promoter [179, 180]. Moreover, owing to these changes in miRNA biogenesis genes, hypoxic tumours are thought to have global reductions of miRNA levels, and these miRNA-deficient tumours with *DICER1* repression have increased propensity towards the epithelial-mesenchymal transition (EMT) and suppressed angiogenesis [180, 181]. Both of these changes have been shown to occur as a result of widespread derepression of miRNA targets in miRNA-deficient tumours. For the case of increased EMT, there is derepression of *ZEB1* through a reduction in miR-200, a known inhibitory factor of the EMT [180]. In the case of suppressed angiogenesis, there is derepression of an inhibitor of the effects of HIF (FIH1), increasing the repression on HIF by FIH1, suppressing angiogenesis [181]. Thus, these tumours undergoing hypoxia-associated changes in miRNA biogenesis experience alterations to their transcriptome, resulting in derepression of miRNA-targeted mRNA, that enable cells to survive this harsh environment, and find alternative ways of selectively producing miRNA in order to adapt effectively.

Recent work by Dr. Laura Winchester and Dr. Simon Wigfield in collaboration with the Buffa and Harris labs, has shown that AGO2 amplification across various tumour types in clinical specimens is a common event co-occurring with hypoxia (manuscript in submission). This effect was shown independently of co-occurring changes to the tumour driver *MYC*, which lies on the same amplicon (8q22-24) as *AGO2*. This study revealed the extensive changes occurring at the level of the miRNA transcriptome in hypoxia, and showed a pan-cancer shift towards AGO2-driven miRNA biogenesis with hypoxia, leading towards the production of oncogenic miRNA. Work from this thesis chapter was incorporated into this manuscript. Moreover, in this chapter I extend this further through a comprehensive analysis of all miRNA biogenesis genes, across 15 epithelial cancer types, with data from the TCGA project.

4.1.5 Research questions

The production of miRNA in the hypoxic microenvironment remains a limitation in the understanding of non-coding RNA biology. In this chapter, I seek to characterise changes at the transcriptomic level occurring in association with hypoxia, and show evidence across multiple tumour types in clinical specimens, that certain miRNA may be preferentially produced through hypoxia-selected AGO2-mediated biogenesis. It is shown how this may lead to enhanced adaptation in the hypoxic microenvironment. That is, I show that there are predicted mRNA targets enabling tumour progression that are negatively correlated to miRNA with reduced expression, and therefore up-regulated in hypoxia, and vice versa. In essence, these observations may help elucidate the evolutionary basis for the selection of this alternative biogenesis pathway.

The remainder of this chapter is structured as follows. In Sections 4.3.1 - 4.3.6, I conduct a pan-cancer analysis of miRNA biogenesis genes, identifying those with recurrent changes in copy number and expression associated with hypoxia. Next, in Sections 4.3.7 - 4.3.9 I study the global dysregulation of miRNA levels and maturation in association with hypoxia, and characterise the miRNA species changing with hypoxia as a result of the biogenesis pathway switch.

4.2 Materials and Methods

4.2.1 Data sources

4.2.1.1 miRNA biogenesis gene panel

A panel of 43 miRNA biogenesis genes, selected from a thorough literature review, as used by Dr. Laura Winchester in her study of miRNA biogenesis machinery genes across tumour types, was analysed. These genes and their functions as pertaining to regulation of miRNA biogenesis are summarised in Table 4.1.

Table 4.1: **Panel of miRNA biogenesis genes considered.** Many show involvement as members of the canonical pathway or mediators of these members.

Gene	Full name	Function
ADAR	Adenosine Deaminase, RNA Specific	RNA editing, conversion of adenosine to inosine, resulting in poorer binding of pri-miRNA to DROSHA [182]
ADARB1	Adenosine Deaminase, RNA Specific B1	RNA editing, conversion of adenosine to inosine, resulting in poorer binding of pri-miRNA to DROSHA [182]
AGO1	Argonaute 1, RISC Catalytic Component	RISC component [183]

AGO2	Argonaute 2, RISC Catalytic Component	RISC component [184]
AGO3	Argonaute 3, RISC Catalytic Component	RISC component [185]
AGO4	Argonaute 4, RISC Catalytic Component	RISC component [185]
ARS2	Serrate, RNA Effector Molecule	Stabilises pri-miRNA in the microprocessor complex [186]
CNOT1	CCR4-NOT Transcription Complex Subunit 1	Scaffolding unit for CCR4-NOT complex, assisting in mRNA degradation by CCR4-NOT complex [187]
DCP2	Decapping mRNA 2	Involved in mRNA degradation for target repression [188]
DDX5	DEAD-Box Helicase 5	Mediate interactions with the microprocessor complex and p53 or SMAD proteins [189, 190]
DDX17	DEAD-Box Helicase 17	Mediate interactions with the microprocessor complex and p53 or SMAD proteins [189, 190]
DDX20	DEAD-Box Helicase 20	Involved in loading miRNA into the RISC, mediates interactions with NF-κB [191]
DGCR8	DGCR8, Microprocessor Complex Subunit	Associated with DROSHA in the microprocessor complex [192]
DHX9	DExH-Box Helicase 9	RNA helicase, interacts with DROSHA in the microprocessor complex to increase pre-miRNA processing [193]
DICER1	Dicer 1, Ribonuclease III	Cleaves pre-miRNA into mature miRNA [194]
DROSHA	Drosha Ribonuclease III	Core component of microprocessor complex [195]
ELAVL1	ELAV Like RNA Binding Protein 1	RNA binding protein that reduces miRNA-target mRNA binding at 3' UTR [196]
ESR1	Estrogen Receptor 1	Interacts with *DDX5/DDX17* (p68/p72) in their modulation of the microprocessor complex [197]
ESR2	Estrogen Receptor 2	Interacts with *DDX5/DDX17* (p68/p72) in their modulation of the microprocessor complex [197]
FXR1	FMR1 Autosomal Homolog 1	Component of RISC, involved closely with AGO2 [41]

GEMIN4	Gem Nuclear Organelle Associated Protein 4	Increases efficacy of RISC for particular miRNA [198]
HNRNPA1	Heterogeneous Nuclear Ribonucleoprotein A1	Binds to DROSHA complex to create more favourable binding in miRNA cleavage site [199]
ILF3	Interleukin Enhancer Binding Factor 3	Forms protein complex with DROSHA microprocessor, overexpression reduces miRNA levels [200]
IPO8	Importin 8	Regulates transport of mature miRNA back into the nucleus [201]
KHSRP	KH-Type Splicing Regulatory Protein	Acts to stabilise AGO2, thereby facilitating miRNA action [202]
MAPKAPK2	Mitogen-Activated Protein Kinase-Activated Protein Kinase 2	Phosphorylates AGO2, enabling localisation to P-bodies and miRNA repression [203]
MOV10	Mov10 RISC RNA Helicase	Interacts with AGO2 to facilitate miRNA repression [204]
PABPC1	Poly(A) Binding Protein Cytoplasmic 1	Interacts with AGO2 to facilitate miRNA repression [205]
PRKRA	Protein Activator Of Interferon Induced Protein Kinase EIF2AK2	Negative regulator of miRNA production, through interaction with DICER [206]
RAN	RAN, Member RAS Oncogene Family	G-protein involved in transport of pre-miRNA through nuclear pore [207]
RBM4	RNA Binding Motif Protein 4	Facilitates miRNA-mediated mRNA repression through interaction with AGO2 [208]
RENT1	Regulator Of Nonsense Transcripts 1	Involved in miRNA degradation through dissociation of miRNA-mRNA binding [209]
SMAD1	SMAD Family Member 1	Interacts with DROSHA to facilitate miRNA production [210]
SMAD3	SMAD Family Member 3	Interacts with DROSHA to facilitate miRNA production [210]
SMAD5	SMAD Family Member 5	Interacts with DROSHA to facilitate miRNA production [210]
SNIP1	Smad Nuclear Interacting Protein 1	Positive regulator of pre-miRNA production through interaction with DROSHA [211]

SRSF1	Serine And Arginine Rich Splicing Factor 1	Splicing factor production of miRNA from introns (miRtrons), produced by DROSHA-independent pathway, and facilitates DROSHA-mediated cleavage [212, 213]
TARBP1	TAR (HIV-1) RNA Binding Protein 1	Interacts with DICER in pre-miRNA cleavage [214]
TARBP2	TARBP2, RISC Loading Complex RNA Binding Subunit	Interacts with DICER in pre-miRNA cleavage [214]
TNRC6A	Trinucleotide Repeat Containing 6A	Assists in localising AGO proteins to P-bodies [215]
TRIM32	Tripartite Motif Containing 32	Enhances miRNA mediated repression efficacy in RISC [216]
TSN	Translin	Assists in pre-miRNA passenger strand degradation in RISC [217]
XPO5	Exportin 5	Exports pre-miRNA from nucleus to cytoplasm [39]

4.2.2 Analytical and statistical methods

4.2.2.1 Analysis of statistical significance of number of miRNA increased or decreased

In doing so, I consider the relationship between the hypoxia gene signature score, and three quantities of total miRNA expression generated by different levels of filtering. Because there are miRNA whose expression remains high and consistent across samples (housekeeping miRNA, for instance), I sought to show how the changes at the global miRNA expression affected different subgroups of miRNA. That is, the changes occurring at the level of all miRNA were characterised, and those miRNA with reduced expression across samples were removed (in order to remove confounding poorly expressed miRNA, which may not vary with hypoxia). Specifically, for this expression filter, miRNA were removed from the analysis if their mean expression across all samples was below the median of all miRNA expressed for a given tissue type. The third subset of miRNA considered was generated by a combination of the above expression filter, and a variance filter. This was designed to ensure that only consider those miRNA which were both expressed and varying were considered, removing any highly expressed, but constant miRNA (such as critical housekeeping miRNA) from confounding the analysis. Specifically, the variance filter removed any miRNA not in the upper quartile of coefficient of variation when considering all miRNA within

a given tissue type. These filters are described algorithmically for further clarity in Algorithms 1 and 2.

Algorithm 1: Algorithm describing the miRNA expression filter employed.

```
function ExpressionFilter (miRNA_names, miRNA_expression);
Input  : miRNA expression matrix
Output: miRNA_passed, miRNA passing expression filter
overall_mean = mean(miRNA_expression)
miRNA_passed = Empty list
for ( miRNA_name in miRNA_names ) {
    if median(miRNA_expression[miRNA_name]) > overall_mean then
        miRNA_passed.add(miRNA_name)
    end
}
return miRNA_passed
```

Algorithm 2: Algorithm describing the miRNA variance filter employed.

```
function VarianceFilter (miRNA_names, miRNA_expression);
Input  : miRNA expression matrix
Output: miRNA_passed, miRNA passing variance filter
for ( miRNA_name in miRNA_names ) {
    coeff_of_var[miRNA_name] =
     mean(miRNA_expression[miRNA_name])/stdev(miRNA_expresssion[miRNA_name])
}
miRNA_passed = which(coeff_of_var[miRNA_name] >
 quantile(coeff_of_var[miRNA_name],0.75))
return miRNA_passed
```

After these three groups of miRNA were obtained, for each miRNA in each group, the question of whether miRNA expression was statistically significantly increased or decreased in samples with high hypoxia score (above the median) versus low hypoxia score (below the median) across all samples in a given tissue type was considered. Expression values were compared using the one-sided Wilcoxon rank-sum test, with a Bonferroni-corrected p value cutoff of 0.05. Across all miRNA considered within a given group, for each tissue type, the proportion of all miRNA which showed either statistically significant decrease or statistically significant increase among the more hypoxic versus less hypoxic tumour samples was calculated.

In determining whether the differences between the number of downregulated miRNA and upregulated miRNA (or their mature:immature ratios) as associated with hypoxia gene expression score, in a given cancer type, were statistically significant, I devised the following approach to calculate significance. To calculate this p value, first note that the distribution of the number of miRNA down and miRNA up follows a multinomial distribution in the most general case, where miRNA down and up refer to the number found to be statistically significant by a one-sided Wilcoxon test. Thus, the probability mass function defining the probability of observing D down miRNA, U up miRNA, given N total miRNA, is given by the multinomial probability density function, where the probability 0.05 is used because this is the cutoff used in the one-sided Wilcoxon rank sum test:

$$\frac{N!}{D!(N-D-U)!U!}0.05^{D}0.90^{(N-D-U)}0.05^{U}$$

Then, the probability of observing a difference greater than or equal to $D-U$ is given by the sum

$$p = \sum_{\substack{d,u\leq N \\ d-u\geq D-U}} \frac{N!}{d!(N-d-u)!u!}0.05^{d}0.90^{(N-d-u)}0.05^{u}$$

which is the desired p value for statistical significance, computed numerically using the above formula for each comparison.

4.2.2.2 Identification of miRNA statistically significantly associated to hypoxic changes

To do this, I devised an ad-hoc statistical approach, summarised pictorially in Figure 4.1.

In brief, the rank (Spearman) correlation is determined for each mature miRNA and the mature:immature ratio of each miRNA to the hypoxia gene signature score, *AGO2* copy number, and *DICER1* copy number. Those miRNA that associate most positively with changes in hypoxia gene signature score and *AGO2* copy number, and most negatively with *DICER1* copy number individually, are collated across cancer types using the rank product statistic, and then the miRNA common to all three groups in this analysis are identified. The results of this analysis are summarised in Figure 4.12, which shows that there are 5 mature miRNA statistically significantly upregulated in association with hypoxia, amplification of *AGO2*, and deletion of *DICER1*. Likewise, on the right panel of Figure 4.12, there are 6 mature miRNA identified as downregulated in association with hypoxia, *AGO2* amplification, and

DICER1 deletion. In the subsequent subsections, these miRNA and their associations are examined.

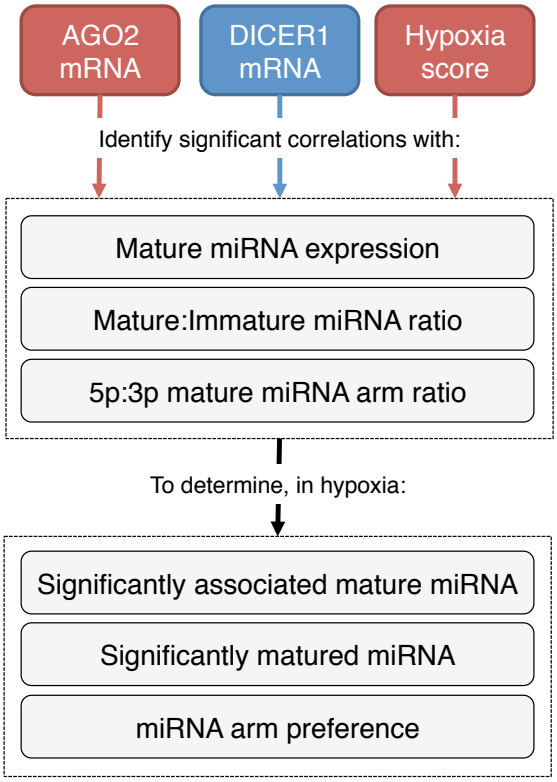

Figure 4.1: **Overview of identification of miRNA associated with hypoxia score increase, *AGO2* increase, and *DICER1* loss.** Red boxes indicate positive statistically significant Spearman correlation coefficients taken, and blue boxes indicate negative statistically significant Spearman correlation coefficients taken. Each of these covariates is correlated with the levels of mature miRNA, mature:immature miRNA ratio, and 5p:3p arm ratio. This identifies, in association with hypoxia (and related changes), the statistically significantly associated mature miRNA, statistically significantly matured miRNA, and miRNA arm preference.

4.3 Results

4.3.1 Copy number and gene expression in miRNA biogenesis genes associates with hypoxia gene signature score

Using matched copy number and gene expression data from the TCGA dataset, as described in the Methods, the statistical association between the copy number of each of the miRNA biogenesis genes and the hypoxia gene signature score across cancer types was analysed. Specifically, for each sample, the hypoxia score as the

median expression of the hypoxia signature genes was computed, and the correlation of this score with the log2-transformed copy numbers (obtained via single nucleotide polymorphism (SNP) array data) and the log2-transformed gene expression level for the 43 miRNA biogenesis genes was determined. Using a similar method as presented in the previous chapter, the rank correlation (Spearman correlation coefficient) was used, with non-statistically significant values reduced to 0, and then applied over the cancer types, resulting in genes with expression and copy number statistically significantly associated both positively and negatively with hypoxia score. In order to weight copy number and expression in equal measure for each cancer type, I considered the rank product statistic on a matrix with two columns for each cancer type, one for the correlation coefficients of copy number with hypoxia score, and the other for the correlation coefficients of gene expression with hypoxia score. This preferentially increased the ranks of those genes correlated with hypoxia score in both domains, and reduced the rank product of those genes not consistently correlated with hypoxia score in both copy number and expression across cancer types.

As summarised in Table 4.2, among the biogenesis genes considered, *AGO2*, along with *PABPC1* and *RAN*, shows the strongest degree of positive association with the hypoxia score. Importantly, among these genes, *AGO2* has been already reported as a key component of the biogenesis machinery amplified in hypoxia in a submitted manuscript by Dr. Laura Winchester, working with the Buffa and Harris labs. In the direction of negative association, there is a very strong signal for the concomitant reduction in *TNRC6*, *DDX5*, *DDX17*, *ESR1*, and two enzymes involved in RNA editing: *ADAR* and *ADARB1*, as hypoxia score increases.

To further understand these changes, this procedure was repeated whilst controlling for the copy number of *AGO2*, as previous work has shown its importance as a potential effector of the hypoxic response. To do this, each correlation coefficient partial to the copy number of *AGO2* was considered, for each of the miRNA biogenesis genes (excluding *AGO2* itself), to adjust for the concurrent changes associated with *AGO2* copy number status and hypoxia gene signature score. It was observed that many of the same genes appeared to be co-deleted or reduced in expression in conjunction with hypoxia gene expression score, even after adjusting for *AGO2* copy number; namely *TNRC6*, *DDX5*, *DDX17*, and *ESR1*, suggesting they are independently statistically associated with the hypoxia gene expression score. Lastly, the *RAN* gene remains statistically significant in its co-amplification status with hypoxia score, again suggesting potential independence as an effector of the hypoxia response for the miRNA biomachinery genes.

Gene name	Copy number/expression change in hypoxia	Rank product p value
TNRC6	Decreased	$2.8 \cdot 10^{-8}$
DDX5	Decreased	$3.3 \cdot 10^{-5}$
DDX17	Decreased	$3.4 \cdot 10^{-5}$
ESR1	Decreased	$1.9 \cdot 10^{-4}$
ADAR	Decreased	$4.6 \cdot 10^{-4}$
ADARB1	Decreased	$5.9 \cdot 10^{-3}$
AGO2	Increased	$7.4 \cdot 10^{-15}$
RAN	Increased	$5.4 \cdot 10^{-14}$
PABPC1	Increased	$4.4 \cdot 10^{-7}$

Table 4.2: **miRNA biogenesis genes showing statistical association in copy number and expression with hypoxia gene signature score.** Table of rank product p value statistic, Bonferroni-corrected for multiple testing, for Spearman correlation coefficients of miRNA biogenesis gene copy number and expression correlated to hypoxia score, across 15 cancer types based on the TCGA dataset.

4.3.2 Association of copy number and expression level

Lastly, I examine the relationship between copy number and expression for each of the miRNA biogenesis genes under consideration. For each tissue type, an empiric distribution for the copy number as correlated to expression for every gene is generated, and then the percentile from this distribution for the correlation of copy number to expression for each biogenesis gene in each cancer type is obtained. Comparing these empirically-obtained percentiles across tissue types for each of the biomachinery genes, as shown in Figure 4.2, demonstrates how copy number and expression correlate across tissue types for each of the miRNA biomachinery genes, with respect to all other genes in the dataset.

As seen in Figure 4.2, for most genes, expression and copy number are correlated above the 50^{th} percentile across most tissue types. This suggests that copy number does play an important role in determining expression for these genes, across most cancer types. However, the genes *ESR1*, *ESR2*, *HNRNPA1*, *ARS2*, *SRSF1*, and *ADARB1* all appear to be correlated poorly (or negatively in some instances) with gene expression across cancer types, suggesting that the dominant mechanism controlling gene expression for these genes is not copy number changes. Conversely, the genes *GEMIN4*, *DROSHA*, *PABPC1*, and *XPO5* all appear to be more statistically significantly correlated in copy number and expression than nearly all other genes, across tissue types, giving evidence to copy number being a very strong determinant of their regulation and expression across cancer types.

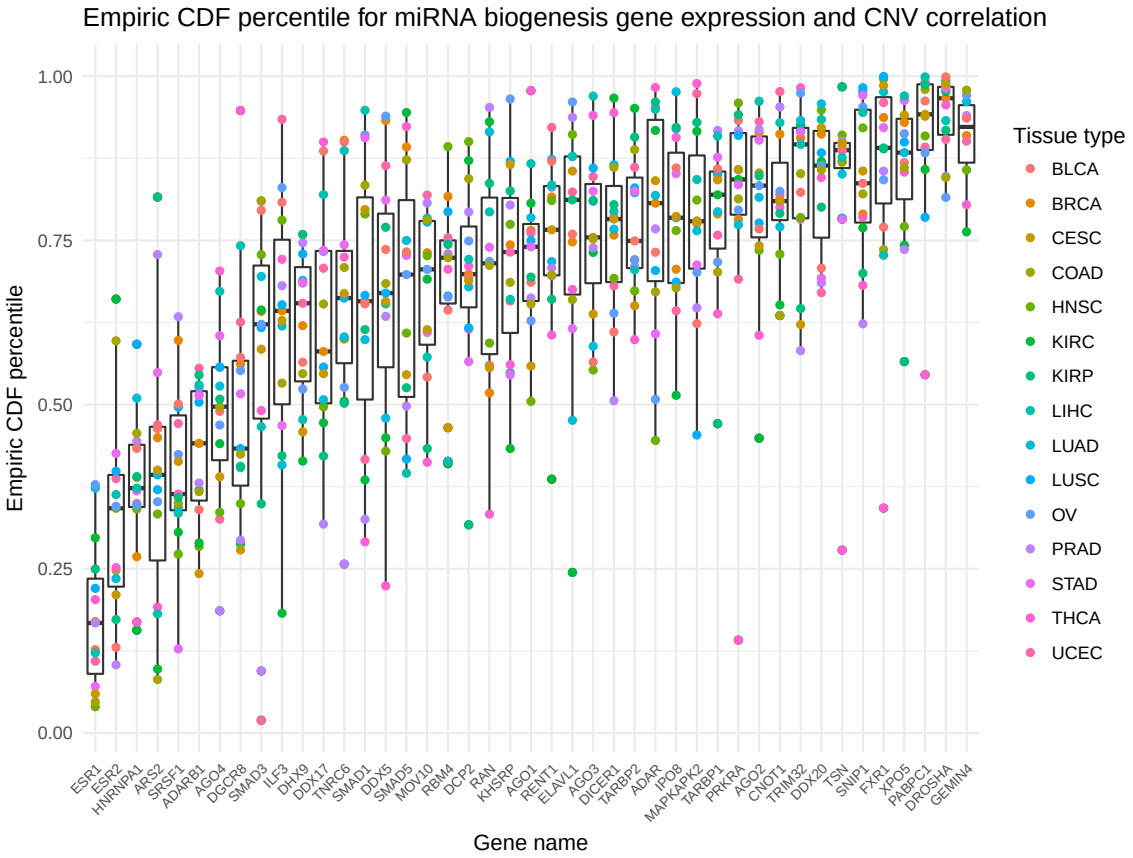

Figure 4.2: **The association of miRNA biogenesis gene copy number and expression.** Empiric percentile for Spearman correlation coefficients between gene expression and copy number for each miRNA biogenesis gene considered. Percentile is computed with respect to the correlation between expression and copy number for all other genes within a given cancer type.

Gene name	Copy number/expr change with *AGO2*	Fold change	p value
TNRC6	Decreased	0.86	$2.5 \cdot 10^{-10}$
DDX17	Decreased	0.89	$6.2 \cdot 10^{-6}$
DDX5	Decreased	0.95	$1.2 \cdot 10^{-3}$
ESR1	Decreased	0.94	$2.3 \cdot 10^{-3}$
RAN	Increased	1.23	$1.2 \cdot 10^{-16}$

Table 4.3: **miRNA biogenesis genes altered in copy number and expression in statistical association with hypoxia gene expression score, adjusted for *AGO2* copy number.** Table of rank product fold change and p value statistic, with Bonferroni correction for multiple testing, taken for Spearman correlation coefficients of miRNA biogenesis gene copy number and expression correlated to hypoxia score, taken partial to *AGO2* copy number, across 15 cancer types based on the TCGA dataset.

Gene name	Copy number/expr change with *AGO2*	Fold change	p value
ESR1	Decreased	0.96	$4.7 \cdot 10^{-8}$
SMAD5	Decreased	0.98	$4.6 \cdot 10^{-5}$
ADARB1	Decreased	1.003	$1.1 \cdot 10^{-4}$
SMAD1	Decreased	0.98	$2.0 \cdot 10^{-4}$
DDX5	Decreased	1.002	$1.4 \cdot 10^{-3}$
ADAR	Decreased	1.05	$6.8 \cdot 10^{-3}$
SMAD3	Decreased	0.999	$7.6 \cdot 10^{-3}$
PABPC1	Increased	1.53	$3.8 \cdot 10^{-44}$
ADAR	Increased	1.05	$9.2 \cdot 10^{-7}$
DHX9	Increased	1.05	$5.2 \cdot 10^{-4}$
RAN	Increased	1.05	$7.1 \cdot 10^{-4}$
XPO5	Increased	1.09	$3.3 \cdot 10^{-3}$
RBM4	Increased	1.06	$6.1 \cdot 10^{-3}$

Table 4.4: **miRNA biogenesis with copy number and expression most correlated to *AGO2* copy number.** Table of rank product fold change and p value statistic (multiple test correction by Bonferroni correction) for Spearman correlation coefficients of miRNA biogenesis gene copy number and expression correlated to *AGO2* copy number, across 15 cancer types based on the TCGA dataset.

Gene name	Copy number/expr change with *AGO2*	Fold change	p value
ESR1	Decreased	0.98	$3.7 \cdot 10^{-7}$
ADARB1	Decreased	1.006	$2.1 \cdot 10^{-5}$
SMAD5	Decreased	0.99	$1.6 \cdot 10^{-4}$
MOV10	Decreased	0.99	$3.3 \cdot 10^{-4}$
SMAD1	Decreased	0.99	$6.6 \cdot 10^{-4}$
ADAR	Decreased	1.05	$2.0 \cdot 10^{-3}$
DDX5	Decreased	1.01	$4.3 \cdot 10^{-3}$
SMAD3	Decreased	1.01	$7.6 \cdot 10^{-3}$
PABPC1	Increased	1.51	$3.4 \cdot 10^{-44}$
ADAR	Increased	1.05	$6.3 \cdot 10^{-7}$
DHX9	Increased	1.05	$2.5 \cdot 10^{-4}$
DDX5	Increased	1.01	$3.1 \cdot 10^{-3}$
TARBP1	Increased	1.01	$5.0 \cdot 10^{-3}$

Table 4.5: **miRNA biogenesis with copy number and expression most correlated to *AGO2* copy number, partial to hypoxia score.** Table of rank product fold change and p value statistic (multiple test correction by Bonferroni correction) for Spearman correlation coefficients of miRNA biogenesis gene copy number and expression correlated to *AGO2* copy number, taken partial to the hypoxia gene expression score, across 15 cancer types based on the TCGA dataset.

4.3.3 Copy number alterations, gene expression, and *AGO2* amplification

As shown, the association of *AGO2* amplification with expression of hypoxia signature genes appears to be a replicable statistical association in clinical samples across tissue types, and one that has been validated experimentally in the recent work by the Harris and Buffa labs (unpublished manuscript). To examine this further, I considered the Spearman correlation of *AGO2* copy number with the copy number and expression of all of the biomachinery genes considered, both independently and adjusted for the hypoxia gene signature score. Using the same methodology of taking non-statistically significant correlations as 0, and then applying the rank product statistic to aggregate results across cancer types, gene lists for biogenesis machinery statistically significantly associated with *AGO2* copy number across cancer types were obtained, before and after adjustment for hypoxia score.

Results of this analysis are summarised in Tables 4.4 and 4.5, showing the unadjusted and hypoxia-adjusted score results, respectively. In brief, these results show that *AGO2* copy number is strongly positively linked with the presence of amplification in *PABPC1*, even more so after controlling for hypoxia, likely because these two genes share the same genomic locus. Further, it is observed that members of the *SMAD* family are decreased in concordance with *AGO2* amplification; namely *SMAD5*, *SMAD3*, and *SMAD1*, even whilst controlling for hypoxia, in addition to *ESR1*. The gene *DDX5*, which shows a decrease overall when compared with hypoxia score, appears to associate statistically significantly both positively and negatively with *AGO2* amplification when controlling for hypoxia, which suggests that its changes may be specific to tissue type.

4.3.4 Co-occurring copy number changes

Given that recurrent copy number alterations in association with hypoxia were identified, I next sought to determine systematically which of these copy number changes co-occurred. The biomachinery genes with consistent positive or consistent negative correlations in copy number with each other were identified, and correlations were adjusted for hypoxia score, across tumour types, thereby identifying scenarios in which co-amplification or co-deletion occurred preferentially, with respect to all other miRNA biogenesis genes, independently of hypoxia gene expression score. This is represented in the heatmap depicted in Figure 4.3, which reveals clusters of strongly correlated biogenesis genes, many of which can be traced back to their originating

genomic loci, as indicated on the heatmap. In addition, it appears that *ADAR* and *DHX9* are associated with copy number gains or losses in most other biogenesis genes, but are positively associated with *AGO2*, suggesting that they may confer a selective advantage when amplified with *AGO2*, but not with other genes.

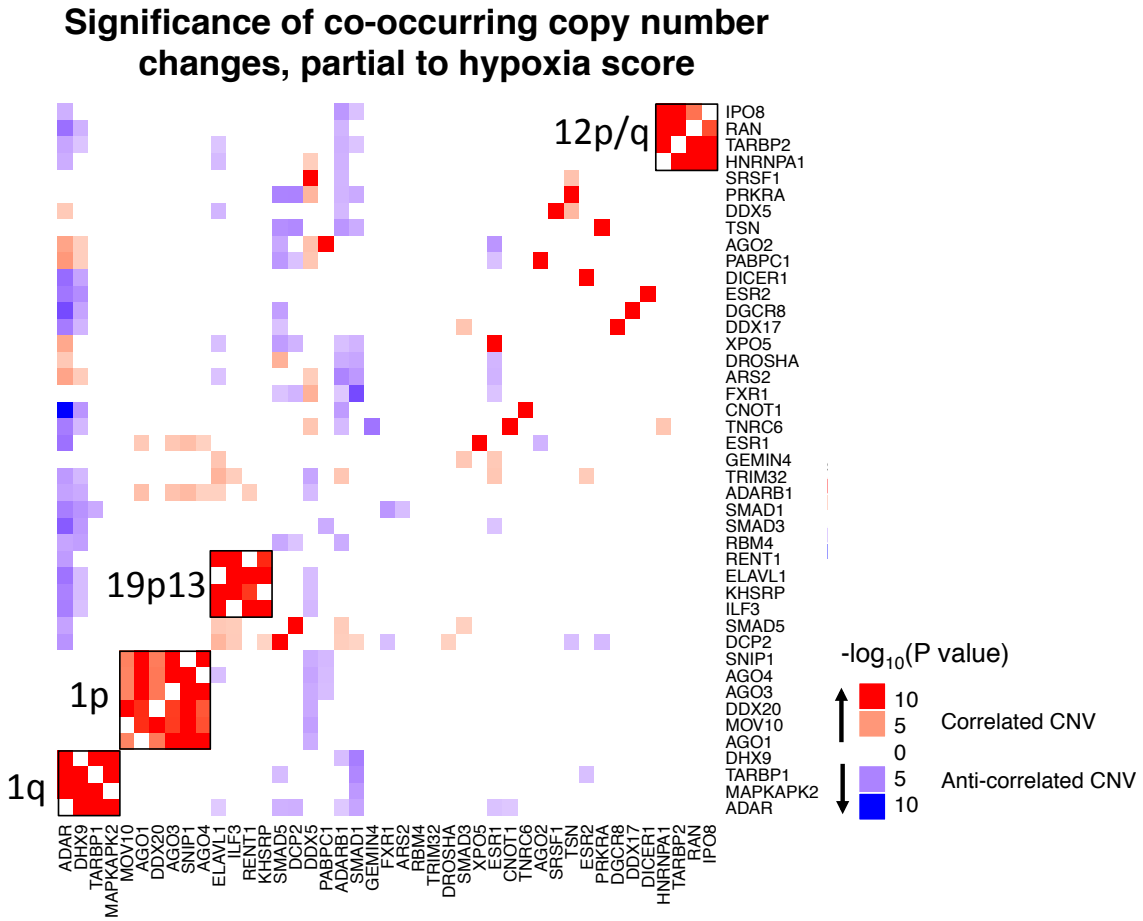

Figure 4.3: **Heatmap depicting statistically significant correlations in copy numbers for miRNA biogenesis genes, partial to hypoxia score, across all cancer types**. Values plotted are the $-\log_{10}$ transform of the p value from the rank product, with a negative value shown if the association between copy numbers is negative (red indicates positive association, blue indicates negative association), and the colour gradient indicates the significance of association. Many statistically significantly co-amplified biogenesis genes share amplicons, as indicated on the heatmap. In the cases where a gene was found to be both statistically significantly positively and negatively associated in copy number with another, the smaller (more statistically significant) rank product statistic was used when indicating significance and directionality.

4.3.5 Mutation frequency and hypoxia

Next, I sought to determine the non-silent mutational frequency for these miRNA biogenesis genes across cancers by using the rank product statistic to identify those increased or decreased in mutation frequency with respect to all other biomachinery genes (Table 4.6). Non-silent mutations for the purposes of this analysis were defined as mutations annotated with the following terms: missense mutation, nonsense mutation, frame shift deletion, frame shift insertion, splice site mutation, translation start site mutation, nonstop mutation, 3' UTR mutation, 5' UTR mutation, 3' flank mutation, 5' flank mutation, in frame deletion, or in frame insertion.

The results from this analysis suggested that *CNOT1* is mutated in greater frequency than all other biomachinery genes, across cancer types. Further, in addition to *CNOT1*, *DICER1*, *DDX5*, and *ADAR*, which are enzymes known to have decreased expression and copy number in hypoxia, were among the most highly mutated genes across samples, suggesting mutation as a further mechanism in reducing their activity. Notably, through this analysis, *AGO2* does not show either a statistically significantly reduced or increased mutation frequency across cancer types with respect to the other biogenesis genes.

This analysis suggested the genes with reduced mutation frequency are *RENT1*, *ADARB1*, and *ELAVL1*. From the previous analysis, *ADARB1* tends towards reduced expression in conjunction with both hypoxia and *AGO2* amplification status, suggesting that a dominant mode of its loss in expression in hypoxia is not due to mutation.

Next, as in the previous analyses, I sought to determine which of the miRNA biogenesis genes were mutated preferentially in concordance with hypoxia score. The mutational status was considered as a binary variable, thereby grouping samples with non-silent mutations against all other samples, and considering the hypoxia score as a response variable. Using a one-sided Wilcoxon rank-sum test, for each biogenesis gene, it was determined whether the hypoxia score was statistically significantly elevated or reduced in samples with non-silent mutations. This procedure was repeated for all biomachinery genes, across all cancer types, and among this resultant matrix of p values from the Wilcox test, the rank product statistic was considered. No statistically significant associations were found between mutational status and hypoxia gene expression score across cancer types (Figure 4.4). Notably, although the changes in hypoxia score among *DICER1* mutated versus unmutated samples is not statistically significant, there is a tendency towards an increased hypoxia score in mutated samples, across multiple cancer types, with similar results for *PABPC1* and *CNOT1*.

In addition to testing the difference in hypoxia score between mutated and unmutated samples for the miRNA biogenesis genes, I also sought to determine whether there were any differences in the copy numbers of *AGO2* and *DICER1* among these samples, as these are also known to be involved in the hypoxic response. Analogous plots to Figure 4.4 for *AGO2* and *DICER1* copy number in place of hypoxia score are shown in Figure 4.5. From the results for *AGO2* copy number, *DICER1*, *PABPC1*, and *CNOT1* showed strong positive association with *AGO2* copy number in mutated samples across multiple cancer types, although not reaching statistical significance. Further, for *DICER1* copy number *CNOT1* appears to show a strong non-statistically significant association with *DICER1* gain, and *ADARB1*, *DDX5*, and *ADAR* show strong non-statistically significant association with *DICER1* loss.

Gene	**Increased/Decreased Mutation freq.**	**Overall mutation freq.**	**p value**
RENT1	Decreased	0%	$5.1 \cdot 10^{-6}$
ADARB1	Decreased	0.7%	$2.6 \cdot 10^{-4}$
ELAVL1	Decreased	0.6%	$2.0 \cdot 10^{-3}$
CNOT1	Increased	2.4%	$5.2 \cdot 10^{-13}$
DHX9	Increased	1.4%	$2.0 \cdot 10^{-6}$
ADAR	Increased	1.2%	$2.6 \cdot 10^{-6}$
DICER1	Increased	1.6%	$1.5 \cdot 10^{-4}$
TARBP1	Increased	1.5%	$9.0 \cdot 10^{-4}$
DDX5	Increased	0.8%	$2.1 \cdot 10^{-3}$

Table 4.6: **Most frequently and infrequently mutated miRNA biogenesis genes, across cancer types.** Table of rank product p values for the frequency of non-silent mutations in miRNA biogenesis genes, ranked across genes, over 15 cancer types, based on TCGA dataset, shown with overall mutation frequencies for corresponding genes, across all tumours considered.

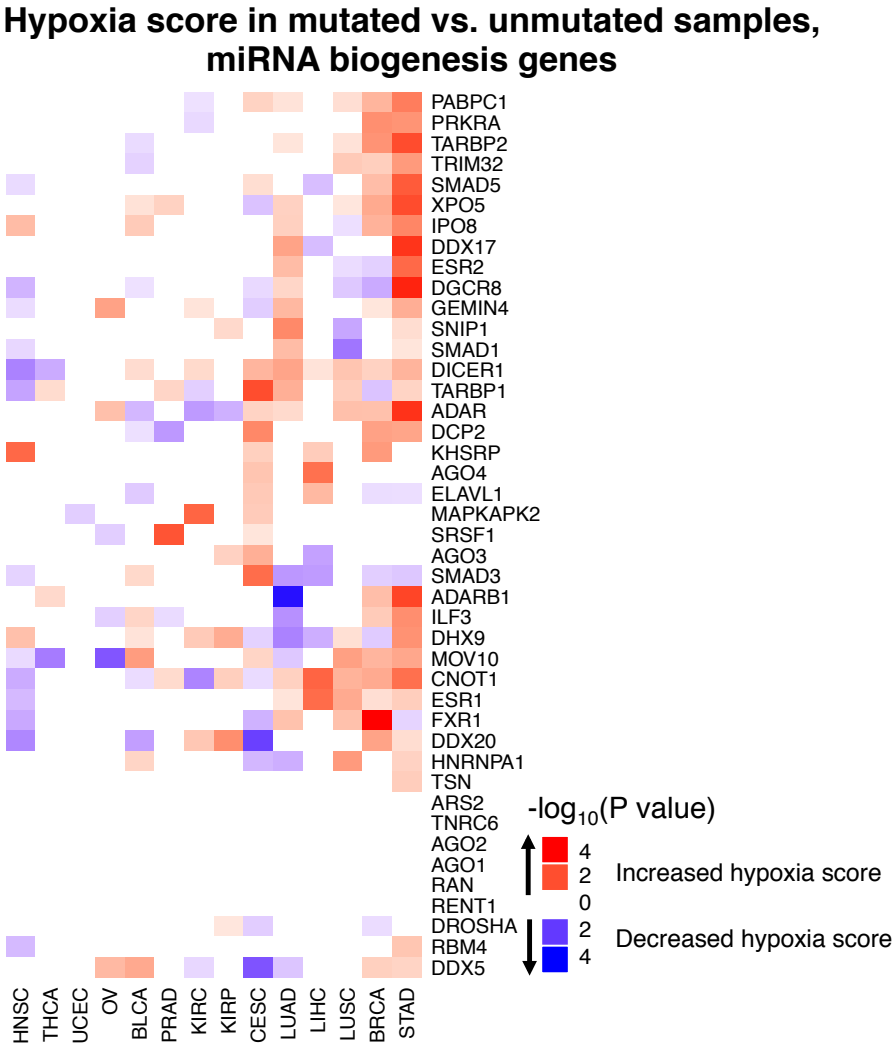

Figure 4.4: **Heatmap showing statistical significance of association of hypoxia score with mutation status for miRNA biogenesis genes, across cancer types.** Mutations considered in this analysis are non-silent mutations. Values depicted show the directionality of the association in copy number (red for positive association, blue for negative association), and the colour scale indicates the significance of the association. That is, blue coloured squares indicate a lower hypoxia score in mutated cases, red indicates a higher hypoxia score in mutated cases, and white squares indicate a case where there were too few mutated samples to test for difference. Statistical difference between the mutated and unmutated groups was determined by a two-sided Wilcoxon rank-sum test.

4.3.6 Genomic alterations to miRNA biogenesis genes co-occurring with hypoxia in breast cancer

I next focus on the changes occurring specifically in breast cancer, particularly with respect to the miRNA biogenesis genes and how they vary with hypoxia. To summarise

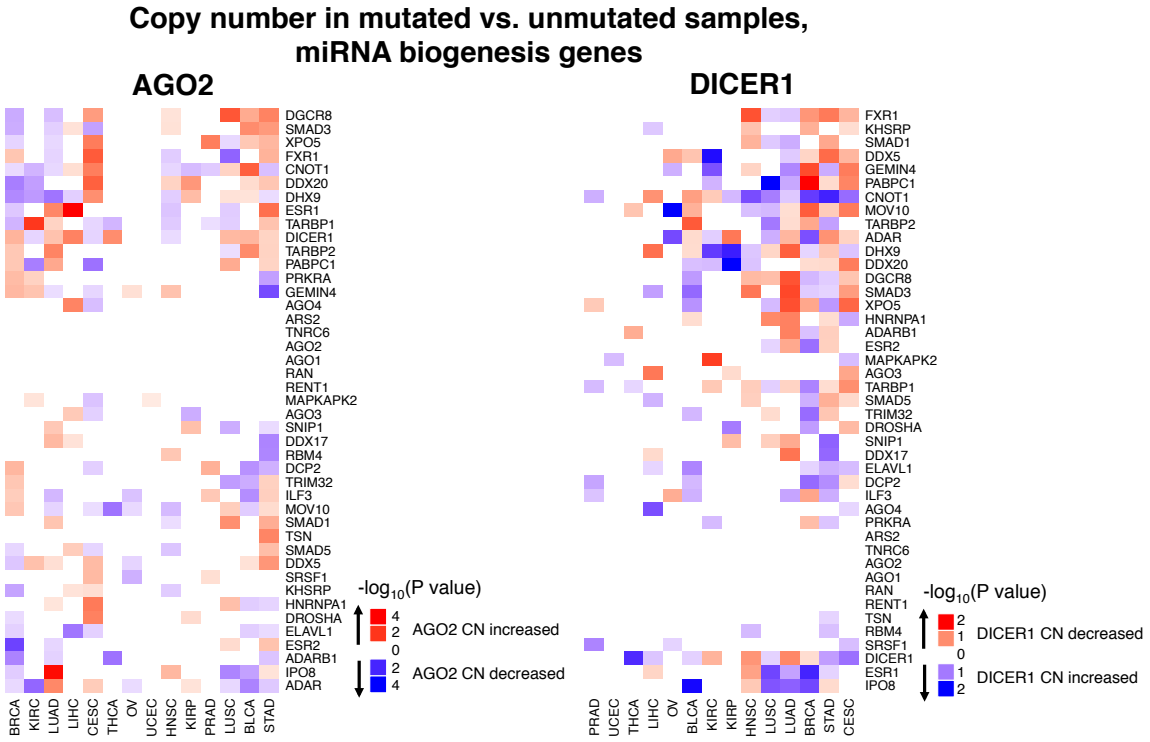

Figure 4.5: **Heatmap showing statistical significance of association of *AGO2* and *DICER1* copy number with mutation status for miRNA biogenesis genes, across cancer types.** Mutations considered in this analysis are non-silent mutations. Values depicted show the directionality of the association in copy number (red for positive association, blue for negative association), and the colour scale indicates the significance of the association. That is, blue coloured squares indicate a lower hypoxia score in mutated cases, red indicates a higher hypoxia score in mutated cases, and white squares indicate a case where there were too few mutated samples to test for difference. Statistical difference between the mutated and unmutated groups was determined by a two-sided Wilcoxon rank-sum test.

the key results from the above sections, it has been shown that across cancer types, *AGO2* is amplified in hypoxia, and *TNRC6* and *DDX5* show evidence for recurrent loss in association with hypoxia. However, because the transcriptomic response to hypoxia may carry tissue-specific effects, the focus was narrowed to the tissue type with the greatest sample size, as this has the largest statistical power to characterise such differences. Thus, in this section, the specific changes co-occurring with hypoxia in the TCGA breast cancer dataset, specifically in cases of invasive ductal carcinoma, are shown.

4.3.6.1 *AGO2* gain and *DICER1* deletion co-occurs in breast and hepatic cancers

Previous work in characterising the changes occurring in the miRNA biogenesis pathway have focussed on breast cancer, and the amplification of *AGO2* and the reduction in expression of *DICER1*, through either copy number changes or epigenetic regulation [180]. Here, changes in copy number and co-expression for these two genes were studied across tumour types. To do this, the Spearman correlation coefficient was computed for the copy number of *AGO2* and *DICER1*, as well as between the mRNA expression values for these genes. Depicted in Figure 4.6, the heatmap shows the values of these correlation coefficients, and as expected based on the literature, there is a slight negative association (i.e. *AGO2* amplification/*DICER1* deletion co-occurrence) in copy number measurable in breast cancer, and a stronger effect at the mRNA level, potentially owing to epigenetic regulation. Interestingly, this effect is also present in hepatic cancers, though to a lesser extent than in breast cancers.

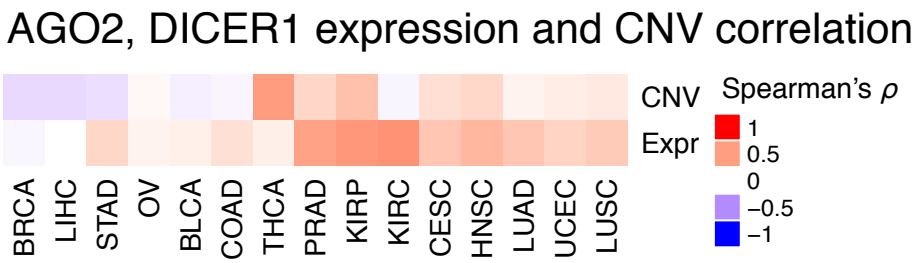

Figure 4.6: ***AGO2* and *DICER1* are inversely correlated in expression and copy number in breast and hepatic cancers.** Heatmap depicting the Spearman correlation coefficients of *AGO2* and *DICER1* copy numbers and mRNA expression values across cancer types.

4.3.7 Alterations to miRNA expression in hypoxic tumours

I next sought to characterise the global changes in miRNA expression across tissue types as a function of hypoxia gene signature score.

As shown in Figure 4.7, the changes in mature miRNA expression are obscured by species that are relatively poorly expressed and not varying, but once these are removed, it becomes clear that for the remaining miRNA, there is a strong trend towards a global decrease in expression as hypoxia gene signature score increases. This result corroborates recent reports of global changes in the levels of miRNA in hypoxic samples, wherein experimentally a reduction has been seen, but this presents

the first evidence of characterising this change at the level of clinical tumour samples, across tissue types.

Next, I sought to determine the association of hypoxia score with the degree of miRNA maturation by analysing the ratio of reads of mature to immature miRNA, across tissue types. I also examined the maturation of miRNA using the same ratio for each of the filtered subsets of miRNA as defined in the previous analysis (Figure 4.7). There was a tendency observed towards global downregulation of miRNA maturation among more hypoxic samples, with particularly large numbers of downregulated mature:immature miRNA ratios in hypoxia for lung, stomach, bladder, and breast tumours. These results are depicted in Figures 4.7-4.11. Note that statistical significance was defined as reported in the Methods section, and is reported as an asterisk above the bars for which the difference between the number of miRNA down versus the number of miRNA up was statistically significant, with significance cutoff $p < 0.05$.

This finding characterises the global expression and maturation changes for miRNA, across clinical samples and tumour types, as a function co-varying with hypoxia. These results corroborate reports of a change in biogenesis in hypoxia, and provides strong evidence for functional differences of miRNA in hypoxic tumours, where *DICER1* tends to be deleted as well.

4.3.7.1 Global miRNA changes and *AGO2* amplification

In addition to the effects co-occurring with hypoxia, as measured through the hypoxia gene signature score, the changes happening at the global level of miRNA expression as *AGO2* amplification occurred were examined. In particular, I studied the proportion of miRNA whose expression (and mature:immature ratio) decreases when *AGO2* is amplified (copy number status ≥ 2.5) as opposed to non-amplified (copy number status < 2.5). In addition, these changes were corroborated with the analogous exercise of studying the effect of increased *AGO2* expression (above or below the median level of *AGO2*) on miRNA expression and maturation. Furthermore, to examine the impact of *AGO2* changes on the different subsets of miRNA; namely all miRNA, those highly expressed, and those that are highly expressed and highly varying, three subgroups for this analysis were again considered (Figures 4.8 and 4.9). This demonstrated that *AGO2* copy number and expression increase are associated with a global trend towards reduced levels of highly varying and expressed mature miRNA. However, the effects of gain of *AGO2* alone were not observed in studying the maturation of miRNA at the global level, suggesting that either there may not

Figure 4.7: **miRNA highly expressed and varying are globally decreased in association with hypoxia.** Graphs describing the proportion of miRNA species showing statistically significant decreases or increases in expression or maturation with hypoxia gene signature score. Statistically significant increase or decrease was computed by the one-sided Wilcoxon rank-sum test between samples above and below the median of hypoxia score for all samples in a given cancer type. Statistical significance in the difference between down and up is depicted by an asterisk over bars for which statistically significantly more miRNA are decreased in number than increased, in a given cancer type. The proportion of miRNA species varying is compared over i) all miRNA, ii) highly expressed miRNA only, and iii) highly expressed and varying miRNA.

be a sufficient number of samples to see an effect, or there are other reasons for why *AGO2* expression or copy number may be increased.

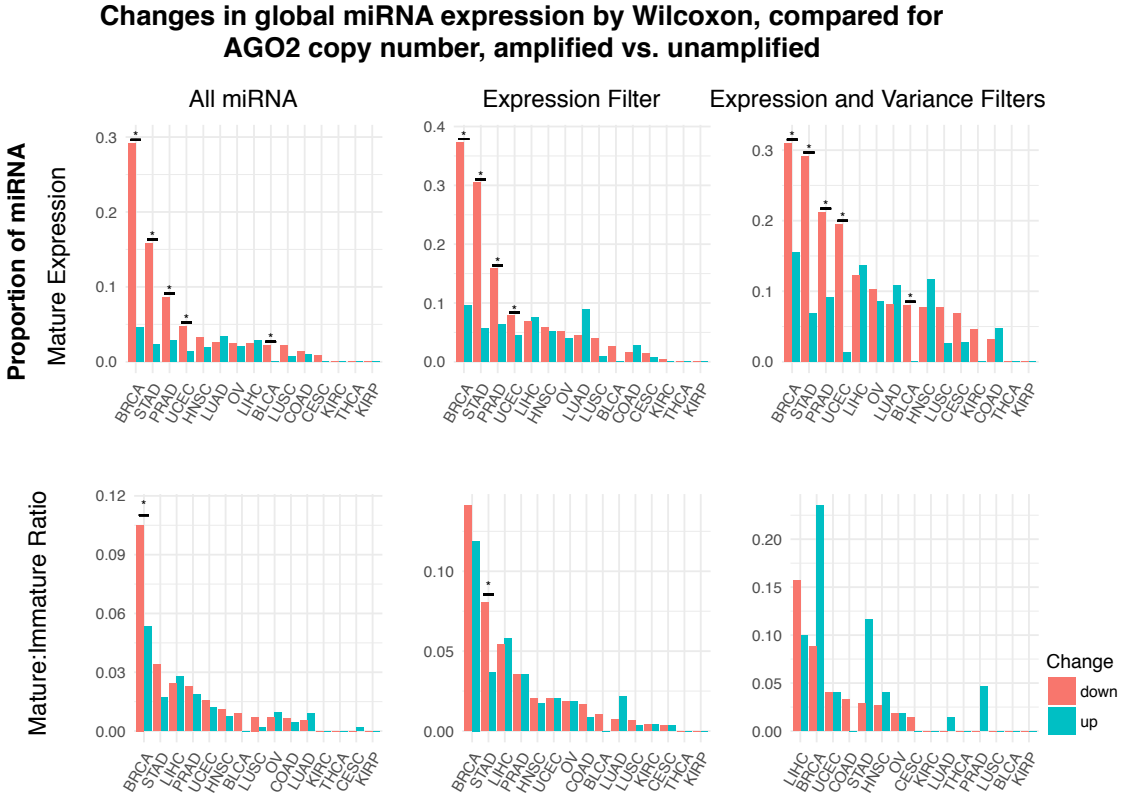

Figure 4.8: **Mature miRNA are globally decreased in association with increases in *AGO2* copy number, but maturation remains unchanged.** Graphs describing the proportion of miRNA species showing statistically significant decreases or increases in expression or maturation with *AGO2* copy number. Statistically significant increase or decrease was computed by the one-sided Wilcoxon rank-sum test between samples above and below *AGO2* copy number of 2.5 in a given cancer type. Statistical significance in the difference between down and up is depicted by an asterisk over bars for which statistically significantly more miRNA are decreased in number than increased, in a given cancer type. The proportion of miRNA species varying is compared over i) all miRNA, ii) highly expressed miRNA only, and iii) highly expressed and varying miRNA.

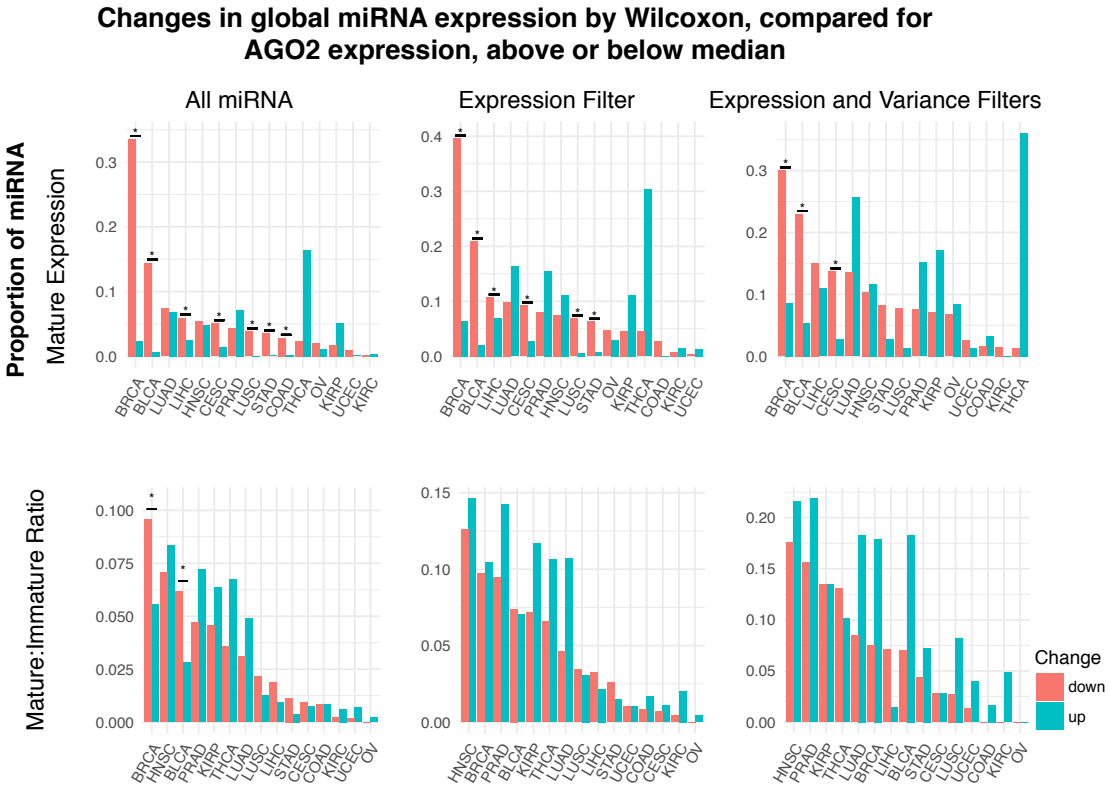

Figure 4.9: **Mature miRNA are generally decreased in association with increased *AGO2* expression, but maturation remains largely unchanged at the global level.** Graphs describing the proportion of miRNA species showing statistically significant decreases or increases in expression or maturation with *AGO2* expression. Statistically significant increase or decrease was computed by the one-sided Wilcoxon rank-sum test between samples above and below the median of *AGO2* expression for all samples in a given cancer type. Statistical significance in the difference between down and up is depicted by an asterisk over bars for which statistically significantly more miRNA are decreased in number than increased, in a given cancer type. The proportion of miRNA species varying is compared over i) all miRNA, ii) highly expressed miRNA miRNA, and iii) highly expressed and varying miRNA.

4.3.7.2 Global miRNA changes in relation to *DICER1* deletion

Similar to the analysis above, I studied the global changes in mature miRNA expression and mature:immature ratio in relation to *DICER1* deletion and reduced *DICER1* expression. Two groups of samples are defined: *DICER1* deleted (copy number ≤ 1.75) or non-deleted *DICER1* (copy number > 1.75), and *DICER1* increased expression (above median expression), or *DICER1* reduced expression (below median expression). Among these groups, the proportion of miRNA statistically significantly up or down for the three subgroups as defined previously (all miRNA, highly expressed miRNA, and expressed and varying miRNA), is compared. The results of this analysis are depicted in Figures 4.10 and 4.11. These results show that *DICER1* deletion is associated with statistically significant decrease in mature miRNA levels across each of the three groups of miRNA considered. Further, as shown in the top row of Figure 4.10, the evidence for a global decrease in mature miRNA levels is particularly evident in breast, bladder, and head and neck cancers. At the level of maturation, particularly for breast cancer, there appears to be a tendency towards the reduction in maturation of miRNA, though this is not statistically significant.

When comparing miRNA levels with respect to *DICER1* expression, as depicted in the top row of Figure 4.11, for the highly expressed and variable miRNA, mature miRNA are preferentially downregulated in breast, renal clear cell carcinomas, renal papillary cell carcinomas, and head and neck cancers. A notable exception is seen for thyroid cancers, where the inverse trend appears to occur. Further, at the level of maturation, among highly expressed and variable miRNA, a statistically significant trend towards downregulation is observed for the miRNA in breast cancer, and this trend is also observed across the three groups for renal cancers and head and neck cancers (bottom row, Figure 4.11).

Figure 4.10: **Mature miRNA are globally decreased in association with decreases in DICER1 copy number, but maturation remains unchanged.** Graphs describing the proportion of miRNA species showing statistically significant decreases or increases in expression or maturation with *DICER1* copy number. Statistically significant increase or decrease was computed by the one-sided Wilcoxon rank-sum test between samples above and below *DICER1* copy number of 1.75 for all samples in a given cancer type. Statistical significance in the difference between down and up is depicted by an asterisk over bars for which statistically significantly more miRNA are decreased in number than increased, in a given cancer type. The proportion of miRNA species varying is compared over i) all miRNA, ii) highly expressed miRNA only, and iii) highly expressed and varying miRNA.

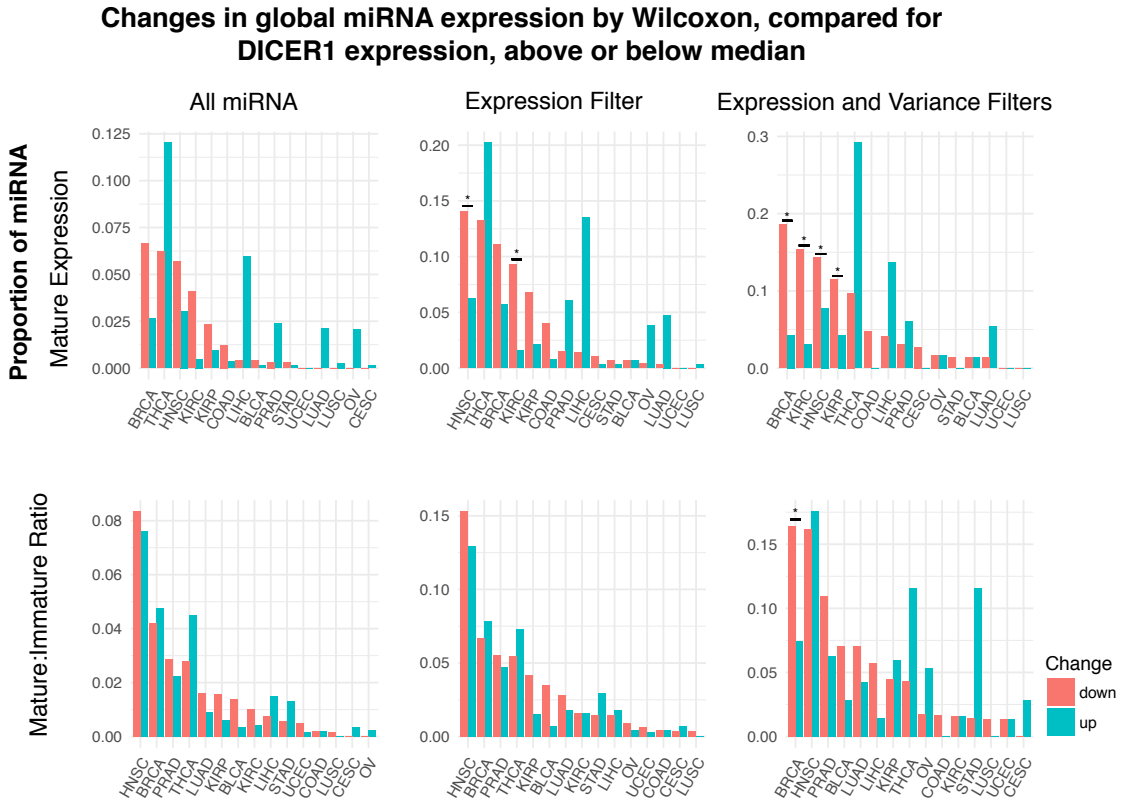

Figure 4.11: **Mature miRNA are generally decreased in association with increased *DICER1* expression, but maturation remains largely unchanged at the global level.** Graphs describing the proportion of miRNA species showing statistically significant decreases or increases in expression or maturation with *DICER1* expression. Statistically significant increase or decrease was computed by the one-sided Wilcoxon rank-sum test between samples above and below the median of *DICER1* expression for all samples in a given cancer type. Statistical significance in the difference between down and up is depicted by an asterisk over bars for which statistically significantly more miRNA are decreased in number than increased, in a given cancer type. The proportion of miRNA species varying is compared over i) all miRNA, ii) highly expressed miRNA only, and iii) highly expressed and varying miRNA.

4.3.8 Associations of miRNA maturation with hypoxia gene expression score, *AGO2* amplification, and *DICER1* deletion

I next determined which specific miRNA contribute most to these observed changes in miRNA levels and maturation, to identify those miRNA most important to the hypoxic response. To reduce the degree of false positives in this analysis, instead of seeking tissue-level changes, I sought to identify pan-cancer signals for changes in the expression of specific miRNA that associate with all three of hypoxia score, *AGO2* copy number, and inversely with *DICER1* copy number.

4.3.8.1 Hypoxia-associated miRNA overlap statistically significantly with miRNA associated with metabolic and proliferative signatures

Using the gene signature-based analysis in Chapter 3, where the miRNA associated with each of 24 gene signatures representative of a number of hallmarks of cancer were identified, I checked whether there was any statistically significant association between the miRNA potentially associated with hypoxic dysregulation of biogenesis and those identified through the analysis done in Chapter 3. In particular, I used a one-sided Fisher's exact test and asked whether the overlap between the miRNA correlated with features consistent with hypoxic dysregulation of miRNA biogenesis and that of Chapter 3 are statistically significant for each of the 24 gene signatures considered.

Intersecting miRNA identified as statistically significantly positively associated in mature form with hypoxia, *AGO2* copy number, and inversely with *DICER1* copy number (Figure 4.12) with the 24 sets of signature-associated miRNA revealed statistically significant overlap with miRNA associated with oxidative phosphorylation ($p = 0.0013$, one-sided Fisher's exact test) and proliferation (Desmedt et al., 2008) ($p = 0.0021$, one-sided Fisher's exact test). Similar analysis of miRNA negatively associated with gene signatures revealed that these hypoxia-associated miRNA are statistically significantly enriched among those negatively associated with the Hallmark: Inflammatory Response ($p = 0.0011$, one-sided Fisher's exact test), and Hallmark: IL2 STAT5 Signalling ($p = 0.0019$, one-sided Fisher's exact test) signatures.

For the 6 miRNA identified as statistically significantly negatively associated with hypoxia score, *AGO2* copy number and positively with *DICER1* copy number, no statistically significant overlaps with signature-associated miRNA identified from the previous chapter were found, using a Bonferroni-corrected p value of 0.05, correcting for multiple testing, against 24 lists of signature-associated miRNA.

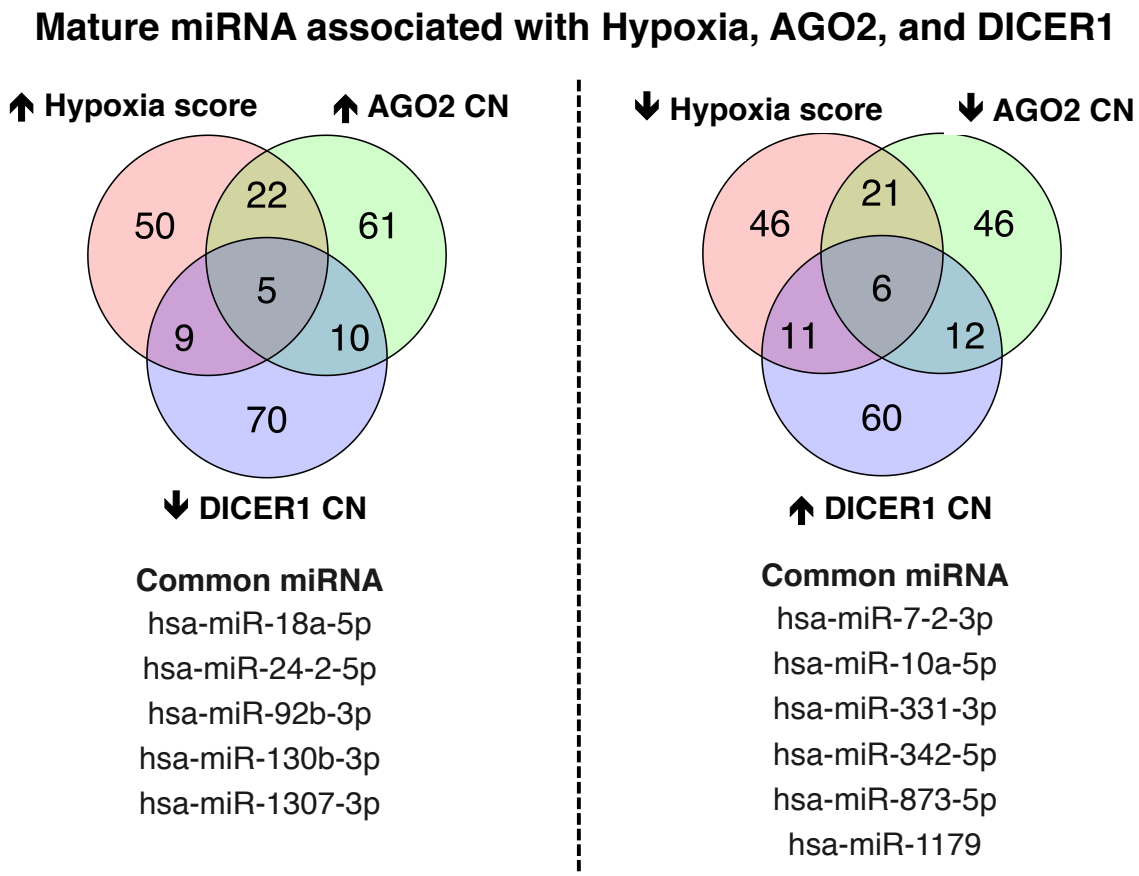

Figure 4.12: **A common set of mature miRNA are statistically significantly positively associated with hypoxia gene signature score, *AGO2* copy number, and *DICER1* copy number, across cancer types.** Left: Venn diagram and listing of common miRNA, showing the overlap of the miRNA statistically significantly positively correlated with hypoxia score, *AGO2* copy number and negatively with *DICER1* copy number. Right: Venn diagram and listing of common miRNA, showing the overlap of the miRNA statistically significantly negatively correlated with hypoxia score, *AGO2* copy number and positively with *DICER1* copy number.

4.3.8.2 Preferentially matured miRNA associated with hypoxia

I next repeated this analysis for the mature:immature ratios of the miRNA in the TCGA dataset, asking which ratios change in conjunction with hypoxia gene signature score, *AGO2* copy number, and *DICER1* copy number. Figure 4.13 depicts the miRNA which are statistically significantly positively and negatively associated in their mature:immature ratio with these miRNA biogenesis alterations co-occurring with hypoxia. As before, to derive an understanding of their potential functions, the one-sided Fisher's exact test was applied to each of these lists of miRNA and it was asked whether the overlap with the signature-associated miRNA lists of the previous chapter was statistically significantly greater than may be expected due to chance.

5 miRNA species were identified as statistically significantly positively associated in mature:immature ratio with hypoxia, and while no statistically significant overlaps with the signature-associated miRNA from the previous chapters were observed, using a stringent Bonferroni-corrected p value cutoff, there was tendency towards association with the Hallmark: Epithelial Mesenchymal Transition signature ($p = 0.0048$, one-sided Fisher's exact test). Further, these 5 miRNA also tended towards overlap, though not statistically significant, with the negatively-associated miRNA for the Invasiveness, Marsan 2014 ($p = 0.0031$, one-sided Fisher's exact test), Angiogenesis, Desmedt 2008 ($p = 0.0034$, one-sided Fisher's exact test), and Hallmark: Hypoxia ($p = 0.0039$, one-sided Fisher's exact test) gene signatures. Thus, these miRNA were shown to positively associate with hypoxia in mature:immature ratio, but tended towards negative association (in mature form) with the hypoxia, angiogenesis, and invasiveness gene signatures. This apparent paradox may be resolved by noting that the known associations that were tested for were in the mature form only. Thus, while hypoxia may decrease the mature forms of these miRNA, here the change in mature:immature ratio was examined, which may increase as a result of a lower production rate of these miRNA, leading to lower immature concentrations of these miRNA, or there may be greater efficiency in maturation (and therefore increased mature:immature ratios) for these miRNA.

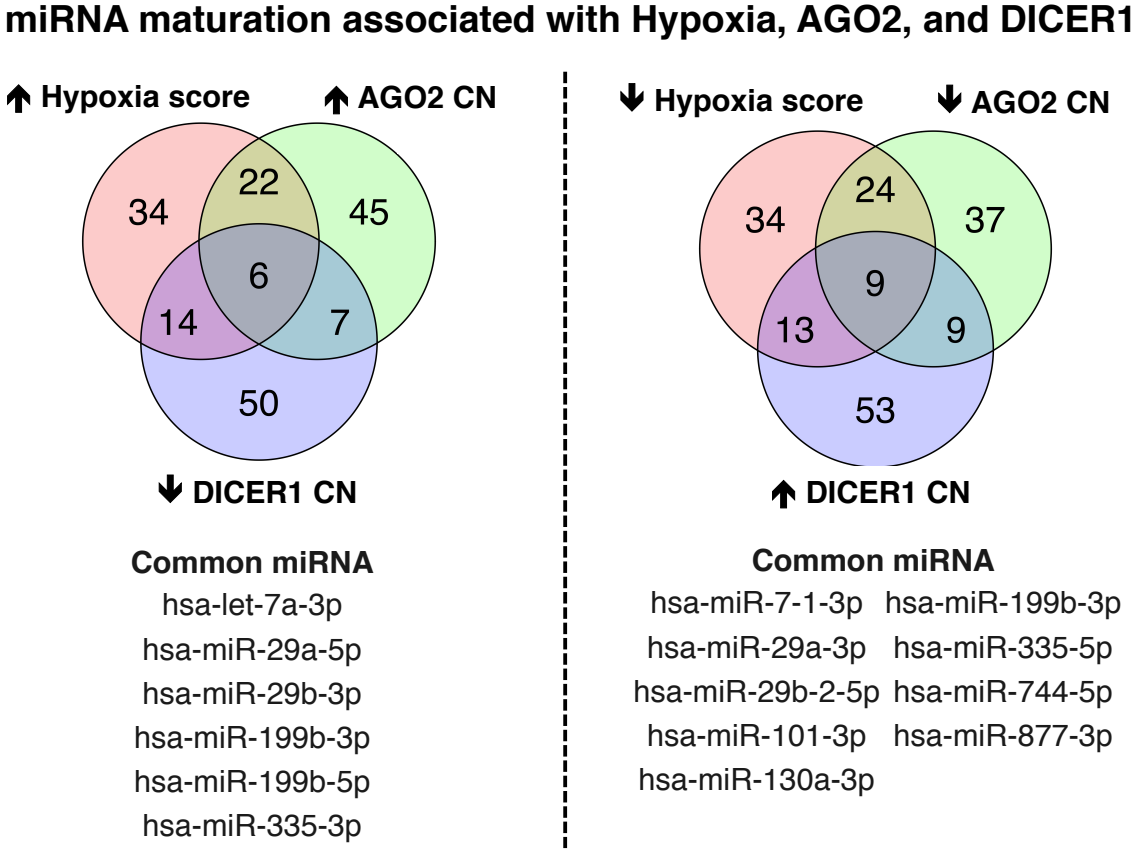

Figure 4.13: **A core set of miRNA are statistically significantly increased in mature:immature ratio in association with hypoxia gene signature score, *AGO2* copy number, and inversely with *DICER1* copy number, across cancer types.** Left: Venn diagram and listing of common miRNA, showing the overlap of the miRNA with mature:immature ratio statistically significantly positively correlated with hypoxia score, *AGO2* copy number and negatively with *DICER1* copy number. Right: Venn diagram and listing of common miRNA, showing the overlap of the miRNA with mature:immature ratio statistically significantly negatively correlated with hypoxia score, *AGO2* copy number and positively with *DICER1* copy number.

4.3.9 miRNA arm selection in hypoxia

Arm selection remains an understudied topic with respect to the regulation of miRNA behaviour. miRNA arm selection has been thought of as a process typically dependent upon binding energies and hydrostatic interactions within the AGO2/DICER1 RISC. However, recent work has shown that miRNA arm selection may be altered in various disease states, with a preference to a particular arm leading to differential repression of targets in some diseases, as has been shown experimentally recently for miR-193a in breast cancers (preferential expression of -3p arm in tumours) [218]. Every pri-miRNA, when processed produces two mature miRNA forms after slicing by DICER1, namely the -3p and -5p forms, and typically these have nearly complementary base pair sequences. The choice of which arm (-3p or -5p) used for the suppressive activity is thought to be primarily due to hydrostatic interactions between AGO2 and the miRNA, with the unselected miRNA arm degraded in the cytoplasm. However, as shown previously, there is a possible change to the miRNA processing machinery in hypoxia, and as this may affect the maturation of miRNA, I sought to identify whether a signal could be detected for the arm selection of miRNA in hypoxic versus less hypoxic samples.

I therefore considered the ratio of the -5p to -3p mature form of the miRNA for each mature miRNA for which both arms were reported in normalised read counts in the TCGA dataset. Considering this ratio across samples, I asked for which miRNA this ratio co-varied the most with hypoxia score, *AGO2* expression, and inversely with *DICER1* expression. This approach yielded a set of miRNA for which, with statistical significance, the -5p to -3p arm ratio covaried with hypoxia and the changes occurring to *AGO2* and *DICER1* in hypoxia. Specifically, as described in Figure 4.14, hsa-miR-29a and miR-199b were identified to have a statistically significantly upregulated 5p:3p ratio and this ratio positively associated with hypoxia gene expression score, *AGO2* copy number, and negatively with *DICER1* copy number. In addition, the hsa-miR-335 5p:3p ratio showed statistically significant decrease in concert with hypoxia, suggesting potential preferential selection for the -3p arm over the -5p arm for this miRNA in hypoxia.

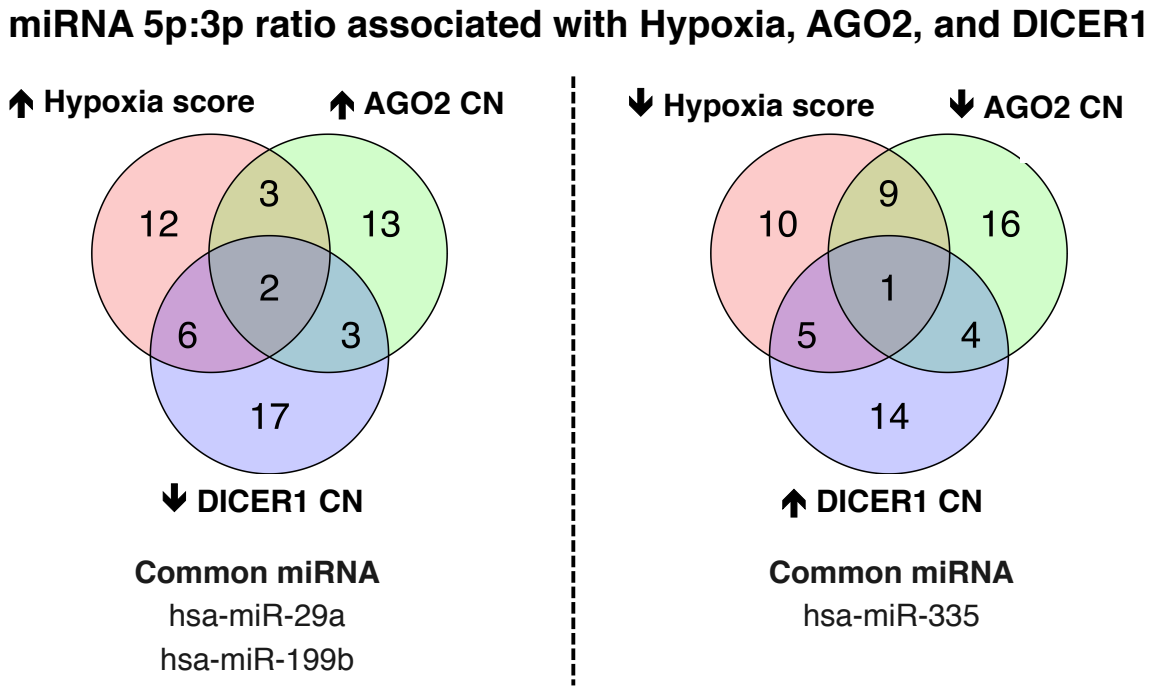

Figure 4.14: **A core set of miRNA show evidence for arm selectivity in statistically significant association with hypoxia gene signature score, *AGO2* copy number, and inversely with *DICER1* copy number, across cancer types.** Left: Venn diagram and listing of common miRNA, showing the overlap of the miRNA with 5p:3p ratio statistically significantly positively correlated with hypoxia score, *AGO2* copy number and negatively with *DICER1* copy number. Right: Venn diagram and listing of common miRNA, showing the overlap of the miRNA with 5p:3p ratio statistically significantly negatively correlated with hypoxia score, *AGO2* copy number and positively with *DICER1* copy number.

4.3.9.1 Differential arm expression in hypoxia correlates with decreased TSG expression and increased oncogene expression

Next, having identified miRNA in which the -5p or -3p arm appeared selected for in hypoxia, I sought to investigate the patterns of association with their predicted targets. Using target prediction in the same manner as done in Chapter 3, a list of predicted targets was generated for the -5p and -3p arms of each of these miRNA, and the rank correlation coefficient between the expression of the miRNA and its predicted target for each miRNA-target pair across cancers was determined. I then took the rank product for the correlation coefficients across cancer types was taken, and this process identified those miRNA-mRNA pairs with correlation coefficients ranked as statistically significantly more negative than due to chance, as compared to all others. These statistically significantly negatively correlated miRNA-target pairs across cancer types are taken as the negatively correlated, and potentially targeted mRNA by each of the miRNA. The list of each of these negatively correlated targets for each of the miRNA is summarised in Table 4.7. For the miRNA statistically significantly positively associated with the hypoxia gene expression signature, and alternative biogenesis pathway genes (miR-29a-5p, miR-199b-5p, and miR-335-3p), there are a number of predicted targets which are tumour suppressor genes negatively associated in expression, and oncogenes positively associated in expression. For instance, among the negatively correlated targets of miR-29a-5p is the *ARID1* tumour suppressor [219], and likewise miR-335-3p is shown to be negatively associated with *CYLD* and *PTEN* [220, 221]. Further, the targets of miR-29a-3p that are increased as a result of its reduction are the oncogenes *TET1*, *PPM1D*, and *BRD3* [222, 223, 224].

miRNA name	-5p targets	-3p targets
hsa-miR-29a	*ARID2, ARL15, ATF2, ATP11B, C3orf58, DENND1B, FAM63B, GOLIM4, MFSD6, PDE7B, PPM1D, PTPRB, RBFOX2, SBNO1, SESN3, SKIL, ZFX*	*ATAD2B, BLMH, BRD3, CBX1, CCDC138, CEP76, COMMD2, CSRNP2, DNMT3A, DNMT3B, ELOVL4, FERMT2, FRAS1, FREM2, GID8, GPAM, GPATCH2, GSTA4, HMGCR, IREB2, KDELC1, KDM5B, KIAA2022, LAMC1, LRP6, LYSMD1, MBTD1, MEX3B, NKAIN1, NKIRAS2, NREP, NUP160, PDIK1L, PPM1D, PRKAB2, PXYLP1, RALGPS1, RIT1, RMND5A, SHPRH, SIKE1, SS18L1, TAF11, TBCCD1, TDG, TET1, TET3, TMTC3, TUBD1, TXNDC16, UBTD2, WASF1, ZBTB5, ZMYM2, ZNF28, ZNF704*
hsa-miR-199b	*AKAP1, AP1G1, ARHGAP12, ARL6IP1, CAPRIN1, CCDC43, CENPI, CLCN3, EIF5B, HAPLN1, HK2, HSPA9, KDM3B, KIAA1958, LARP4, LCLAT1, LIN7C, MARCH8, MPP5, NUDT5, PNPT1, PRPF40A, RAB10, RAD23B, RANBP2, RBBP4, RBM47, SOS2, TMPO, UNG, USP19, WDR76, YIPF6, ZNF440, ZNF709*	*ACVR2B, CD2AP, CDK7, CELSR2, CHKA, DEPDC1B, ESRP1, ETNK1, FAM199X, FAM60A, G3BP2, KATNBL1, KIAA0319L, KTN1, LLGL2, LRP2, LRRC1, MARC1, NAA25, NLK, PAK4, PLEKHH1, PPP1R9A, RBM47, SLC22A5, SMIM8, TFAM, THAP9, TMEM161B, TRMT61B, YWHAE*
hsa-miR-335	*ADGRL2, ALOX5AP, ARF4, ATP2B4, CASP7, CNN1, DAAM1, FAM107B, FAS, GBP1, HOXD8, ITGB2, KRT24, PLOD1, PTPN22, RBFOX2, SGMS2, ZEB2, ZMPSTE24, ZSWIM8*	*AMPD3, ANTXR2, ANXA5, ARL15, ATP6V1B2, BBOX1, CCDC170, CCNDBP1, CCPG1, CD55, CD82, CHD9, CLIC2, CNTN1, COL8A1, CPEB4, CPXM2, CRISPLD2, CYBRD1, CYLD, DPYD, EDNRA, EIF4E3, ENTPD1, ERAP1, ERMN, FAM19A5, FAM63B, FGL2, GOLGA2, GPR155, IGF1, IL16, IL7R, ILK, ITGAV, ITGBL1, KCNMA1, LMO4, LRRC8C, LYPLAL1, MEF2C, NCKAP1L, NECAB1, NEXN, NFIA, NPR3, NRIP1, PDLIM5, PER3, PMEPA1, PRRX1, PTEN, PTGER3, PTPRC, RORA, RUNX1T1, SFMBT2, SGMS2, SPARC, SPATA18, SWAP70, SYNPO2, TAGAP, THBS2, TSHZ3, VWA5A, WDR7, ZCCHC24, ZCCHC5*

Table 4.7: **Predicted targets of miRNA arms upregulated in hypoxia are statistically significantly negatively correlated with many tumour suppressor genes.** Predicted miRNA targets showing statistically significant negative correlation across cancer types, for each of the miRNA identified as associated with arm selectivity in hypoxia, organised by -5p and -3p predicted targets.

4.4 Discussion

In this chapter, through an analysis of the changes occurring to a panel of miRNA biogenesis genes, the alterations most associated with tumour hypoxia were identified. The work presented above showed that such changes frequently involve amplification of *AGO2*, and in some cases, deletion of *DICER1*. With the added insights from previous work characterising non-canonical miRNA biogenesis pathways, it was hypothesised that an alternative biogenesis pathway involving *AGO2* preferentially over *DICER1* may operate in hypoxic tumours. It was shown that the implications of this hypothesis have the potential to explain many of the changes observed in hypoxic tumours, through preferential maturation of specific miRNA involved with the EMT and reducing inflammation. The results within this chapter may suggest a coordinated response to tumour hypoxia, involving multiple species of non-coding RNA working in tandem to adapt to the selective pressure conferred by tumour hypoxia. In doing so, it is posed that tumours are able to achieve changes enabling their proliferation even within a hypoxic niche of tissue, and as has been substantiated by multiple lines of experimental evidence, these cells are those which disproportionately contribute to invasiveness, metastasis, and poor prognosis.

4.4.1 Alterations to certain miRNA biogenesis genes are consistently associated with hypoxia gene signature expression

As discussed above, the canonical miRNA biogenesis pathway is altered by the presence of cellular stressors, signal transducers, and microenvironmental conditions, and in this chapter I have added to this understanding by uncovering relationships of miRNA biogenesis genes to hypoxia using a hypoxia gene signature [174, 180, 189]. As a previous study from the Buffa and Harris labs has shown, and replicated in these findings, hypoxia associates with an increase in both the copy number and expression of *AGO2* and *PABPC1* across cancer types, potentially due to a switch to AGO2-dependent, DICER1-independent biogenesis (unpublished manuscript). Here, this is extended, and it is shown that *TNRC6*, *DDX5*, and *DDX17* deletion and reduced expression are common, statistically significant, co-occurring events in conjunction with hypoxia, across 15 epithelial tumour types.

The loss of *TNRC6*, statistically significantly associated with hypoxia and *AGO2* amplification, has been reported previously in the literature in an analysis of mutations of *AGO2* and *TNRC6* in colorectal cancers [225]. *TNRC6* is a key organisational

component within the P-body, promoting the efficient miRNA-mediated mRNA repression [226]. Thus, its loss in hypoxia may underscore part of a change towards an overall less efficient, but more selective miRNA-mediated repression across the transcriptome.

Further, there is recurrent statistically significant loss of the RNA helicases *DDX5* and *DDX17* in association with hypoxia even when adjusting for *AGO2* status, across cancer types. These RNA helicases serve as cofactors involved in the microprocessor complex through interactions with DROSHA, and transmit signals from p53 to this complex [189, 227]. Under normal circumstances, p53 interacts with DDX5 and DDX17 to enable the more efficient production of miRNAs involved in growth suppression to further effect its protective role in the cell [189, 227]. However, this interaction has only been shown to occur in the case of wildtype p53; this function is no longer critical in p53 mutant tumours, as most hypoxic tumours are, and so loss of *DDX5* and *DDX17* may further perpetuate this, by causing less efficient production of miRNAs involved in protective roles against cancer [189, 227]. Moreover, the RNA helicase ability of these enzymes has been shown to selectively increase the loading of the tumour suppressive miRNA let-7 into the RISC [228, 229]. Thus, the loss of these enzymes likely facilitates tumour progression directly though less efficient mRNA repression of the let-7 targets [229].

Based on these observations, as previous reports confirm, the miRNA biogenesis pathway is one that is highly plastic, depending on the cellular state, and even modifications of the canonical pathway may result in disruption to the miRNA transcriptome.

4.4.2 Preferentially matured miRNA in hypoxia are potentially associated with metabolic changes and inflammation

In addition to the potential modification of the canonical pathway and its regulators, as discussed above, one of the strongest signals across tumour types in association with hypoxia is the amplification of *AGO2*. Given this finding, along with previous reports on the subject, and studies relating hypoxia to epigenetic and copy number mediated *DICER1* loss across cancers, it was hypothesised that hypoxia may mediate a preference away from DICER1-dependent biogenesis and towards AGO2-mediated biogenesis. The existence of this alternative pathway has been substantiated by knockdown experiments performed in [176], where cell lines expressing *DICER1*

knockdown showed that the slicing function could be taken up by AGO2 for particular miRNA. However, because this function only applies to particular miRNA, a preference towards DICER1-independent biogenesis, would create a relative deficit of matured miRNA, particularly for those miRNA that vary highly and may not be necessary for basic cellular functions such as housekeeping miRNA. However, these global changes in the miRNA transcriptome may actually be beneficial in the hypoxic state, as *DICER1* loss and *AGO2* gain may be the products of natural selection in hypoxia, and this change to the miRNA transcriptome may itself be the primary advantage arising from this selection.

To substantiate this, a global reduction in the maturation of miRNA in hypoxia, *AGO2* amplified, and *DICER1* deleted clinical cases was shown, in Figures 4.7- 4.11, which show that across tumour types, there is a generalised reduction in the mature form for most, but not all miRNA. Further, the miRNA that increase statistically significantly in maturation with hypoxia are those termed preferentially matured in the hypoxic milieu. These preferentially matured miRNA may act as a key survival or adaptation mechanism by the cell in order to withstand the stress of hypoxia. Using the statistically significant signature-associated miRNA associated with the hallmarks of cancer from Chapter 3, it was shown that these miRNA were enriched among those associated positively with metabolic changes, such as oxidative phosphorylation and cellular proliferation. It was also shown that these miRNA were enriched among those negatively associated with an inflammatory response gene signature and an IL2-STAT5 signalling gene signature.

Interestingly, the miRNA role in hypoxia is thought to focus in part on changes in metabolism, and specifically the induction of oxidative phosphorylation. Indeed, because miRNA are some of the widest-acting regulators of the transcriptome, they are the ideal candidates to be involved in processes associated with metabolic and inflammatory change. For instance, the well-known hypoxia miRNA, miR-210 is involved in orchestrating the necessary changes associated with altered metabolism in hypoxia, through repression of the iron-sulfur cluster assembly proteins ISCU1/2 [230]. In this analysis, although miR-210 did not show negative association with *DICER1*, it was statistically significantly positively associated with hypoxia and *AGO2* copy number in mature form.

Moreover, the miRNA that tended to associate with hypoxia were those that were potentially negatively involved with inflammation; Wu et al. also commented on this in a recent work, wherein they show that hypoxia-responsive miRNAs function to

repress the cytosolic DNA damage sensor, reducing the triggering of the inflammatory response in DNA damage [231]. These findings have also been experimentally observed in conjunction with tumour hypoxia in previous studies [232]. The overall effects of the changes in the miRNA transcriptome associated with these alterations in biogenesis genes are summarised graphically in Figure 4.15.

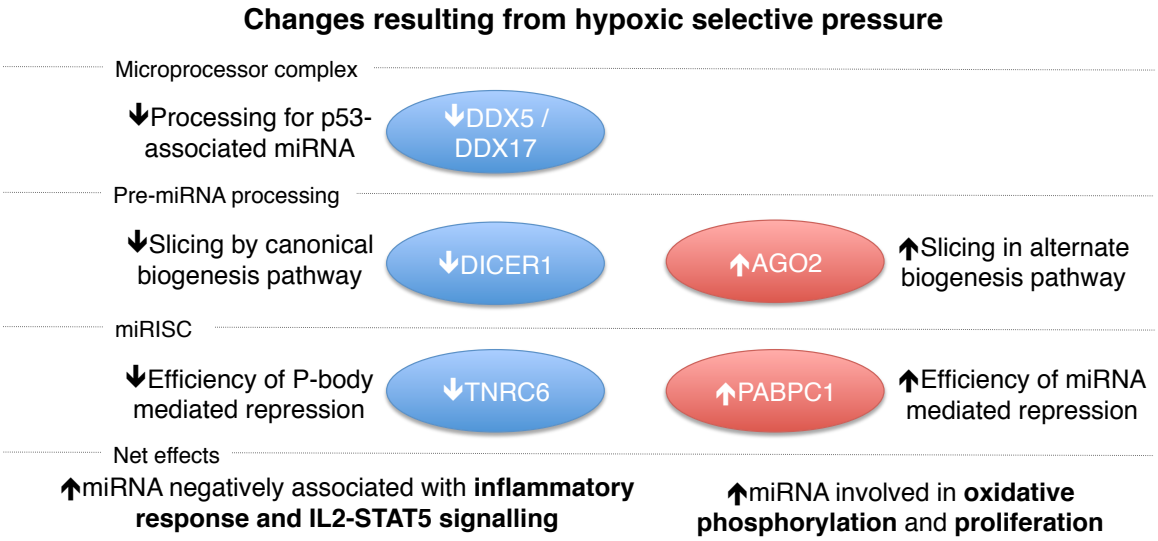

Figure 4.15: **Summary of miRNA maturation biogenesis gene changes and the resulting net effects.** Graphical depiction of the biogenesis genes that are associated with increased copy number and expression in association with hypoxia, and known hypoxic changes, such as increased *AGO2* and decreased *DICER1* expression. The stage of miRNA biogenesis at which each gene has an effect is indicated, as well as the immediate effect on processed miRNA. The resulting changes to the miRNA transcriptome result in increased miRNA related to proliferation, oxidative phosphorylation, and a reduced inflammatory response.

4.4.3 miRNA arm expression correlates with alterations in miRNA biogenesis genes

The selectivity of different miRNA arms (i.e. -5p vs -3p) being produced in different cell types remains an understudied area, and is certainly relevant to biology, as these sequences are nearly complementary and so the two arms of the same miRNA will repress very different targets. Current thinking about miRNA arm selection is based around the hypothesis that the arm preferentially used in repression for a given miRNA is based on thermodynamic interactions with the DICER1 enzyme, during the slicing step, and the opposite arm is degraded by cellular RNAse enzymes [233]. However, in the setting of *DICER1* loss, as occurs in hypoxia, because it is hypothesised that AGO2 compensates for DICER1 loss by contributing to the slicing function, ratios of arm selection of miRNA may change. In this chapter, I investigated the effect of this across cancers, using a methodology combining the changes happening concurrently with hypoxia, *DICER1* loss, and *AGO2* gain, to reduce the potential for false positives. These results, as summarised in Figure 4.14, show that miRNA arm selection does indeed change for specific miRNA in conjunction with hypoxia, *AGO2*, and *DICER1*, and these preferentially-selected miRNA may have profound downstream implications for the cell. Previous reports have studied the 5p/3p arm ratio for various miRNA in different tissue types, and have proposed these as potential biomarkers for disease status and site, as the ratio of miRNA produced, but do not link these to changing biogenesis pathways [234, 235].

An extreme case of *DICER1* loss, namely siRNA-based knockout, has been studied in detail by Kim et al. in [176]. In this work, it was shown that i) the miRNA transcriptome changed in a large way in the case of *DICER1* knockout, and ii) there was a global decrease in miRNA in this setting. Moreover, Kim et al. comment on the observation that in the setting of *DICER1* knockout, the reduction of -3p miRNA, in a general sense, is more severe than the reduction of -5p miRNA, suggesting differential arm selection in this case. However, this work does not go on to characterise the effects of this change in arm selectivity at the level of the targets, leaving the functional implications of these changes unknown.

The miRNA for which the 5p:3p arm ratio correlated to the greatest degree in conjunction with the changes in hypoxia across cancer types were identified; namely, miR-29a, -199b, and -335. These are well-known tumourigenic miRNA, again suggesting that the dysregulation that occurs with the change in biogenesis pathway functions to promotes cancer progression. In fact, miR-335-3p has been validated as involved in cell proliferation, differentiation, and migration across cancer types, and

has been proposed as a prognostic biomarker for gastric cancers [236, 237, 238, 239]. Similarly, miR-29a-5p has been proposed as a part of biomarkers for colorectal, gastric, and hepatocellular cancers [240, 241, 242], and miR-199b-5p has been proposed as a biomarker in endometrioid carcinomas [243]. Lastly, as discussed in the results, among the targets of these miRNA whose -5p or -3p arms appear to be preferentially selected, there is evidence for pan-cancer repression of tumour suppressor genes, which itself may further contribute to cancer progression.

Thus, the miRNA that associate statistically significantly with changes in arm selectivity, hypoxia, and the changes associated with a preference in miRNA biogenesis pathway, may have biological backing in their importance. Given the evidence for a preference towards differential utilisation of biomachinery proteins from DICER1 to AGO2-mediated slicing, and because DICER1 is theoretically responsible for arm selection, one of the first signs of this preference in miRNA biogenesis may be a differentially expressed 5p:3p arm ratio. Further, as the above discussion has shown, these specific miRNA arms have already been identified as driving cancer progression, and have been identified as biomarkers, potentially for this underlying reason; they capture the functional change occurring with a switch in miRNA biogenesis.

4.5 Summary and Conclusions

In this chapter, I have shown how, through an analysis of miRNA biogenesis genes across 15 cancer types, in conjunction with the hypoxia gene signature score, the changes in the miRNA biogenesis pathway may be identified. It was shown that the major changes to the biogenesis pathway involve the deletion of *TNRC6*, *DDX5* , and *DDX17*, the amplification of *AGO2* and *PABPC1*, and reduced expression of *DICER1* in breast cancers. It was then shown how these changes, potentially indicative of an alternative miRNA biogenesis pathway in hypoxia, change global miRNA maturation patterns, and how selectively matured miRNA contribute to adaptation in the hypoxic milieu. Overall, these results suggest that there is indeed global miRNA dysregulation in hypoxia, and it may arise because of a selective switch towards an alternative biogenesis pathway, enabling the more efficient production of miRNA involved in tumour progression and evading the immune response.

Chapter 5

A circRNA antisense to *HSP90AB1* may potentiate the switch to DICER1-independent miRNA biogenesis in hypoxic breast cancers

Abstract

In this chapter, I explore the changes occurring to the circRNA transcriptome in breast tumours in statistical association with hypoxia gene signature score. I uncover evidence for the association of a circRNA antisense to the *HSP90AB1* gene with the hypoxia gene signature, with this circRNA showing predictive ability for hypoxia gene signature score and positive association with *AGO2*. This circRNA also showed strong, statistically significant, correlation with its linear sense transcript, *HSP90AB1*, which may be indicative of stabilisation, and this transcript itself had expression correlated with *AGO2*. This suggested that this circRNA may be involved in a mechanism facilitating the increased expression of *AGO2* in hypoxia, associated with the hypothesised change in miRNA biogenesis previously identified. I also showed that this circRNA was detectable among two of four MCF-7 cell line samples, and in those detectable, did increase in the hypoxic condition. Additionally, in this chapter, I examined the circRNA profiled using three different RNA-seq preparations, and two different circRNA computational pipelines for the MCF-7 cell line after exposure to hypoxia and normoxia. I characterised the similarities and differences between the circRNA identified from these differing preparations and pipelines, and through this provided guidance on future experimental protocols to study circRNA and the hypoxic response.

5.1 Introduction

5.1.1 circRNA biogenesis and potential functions

Circular RNA molecules, denoted circRNA, are recently identified transcripts of RNA found to be circularised within cells [29]. Originally thought to be splicing byproducts, these circularised transcripts are conserved across multiple species, and are created via a well-defined biogenesis pathway [244]. circRNA are transcribed from a pre-mRNA gene in the genome by RNA polymerase II, and then undergo a process known as back-splicing wherein the 5' end of the RNA molecule is covalently linked to the 3' end, resulting in a circularised RNA molecule [245]. The process of back-splicing is mediated by spliceosomal machinery, and multiple reports have suggested that the production of circRNAs is linked inversely with the generation of splice variants, through competition for spliceosomal machinery [245, 246]. The formation of circRNA involving exons from the pre-mRNA transcript is thought to occur through two primary mechanisms: direct backsplicing, and exon skipping followed by intra-lariat backsplicing [245]. Direct backsplicing refers to two exons being joined from the 3' tail to 5' head directly (Figure 5.1A). Exon skipping occurs when two or more intervening exons are removed (or skipped) while a linear splice variant is produced, and these removed exons within the lariat structure undergo back-splicing to circularise, as shown in Figure 5.1B, adapted from [245]. Intronic circRNAs are thought to be derived from lariat RNA more directly, and like lariat RNA, feature a 2'-5' covalent bond, as opposed to the 3'-5' bond found in exonic circRNAs [245].

As a result of their circular conformation, circRNA are relatively insensitive to traditional forms of RNA degradation within the cell, and have been shown to exhibit increased stability as a result [247]. This unique property of circRNA enables them to achieve potentially highly robust and stable function over prolonged cellular timescales. In fact, circRNA are thought to accumulate within cells as they age, and therefore reach differing concentrations as a function of the division rate of the cell type they are present in [245]. Many genes expressing splice variants have shown concomitant circRNA expression at detectable levels, but only approximately 50 genes show circRNA expression at levels higher than linear variants of the gene [248]. As such, circRNA were originally thought to be byproducts of alternative splicing, but functional roles for certain circRNA are now emerging. For instance, there is ample evidence for highly tissue specific expression, such as high expression in mammalian brains [244, 249], involvement in neural developmental processes [250, 251], and involvement in processes such as the EMT [252].

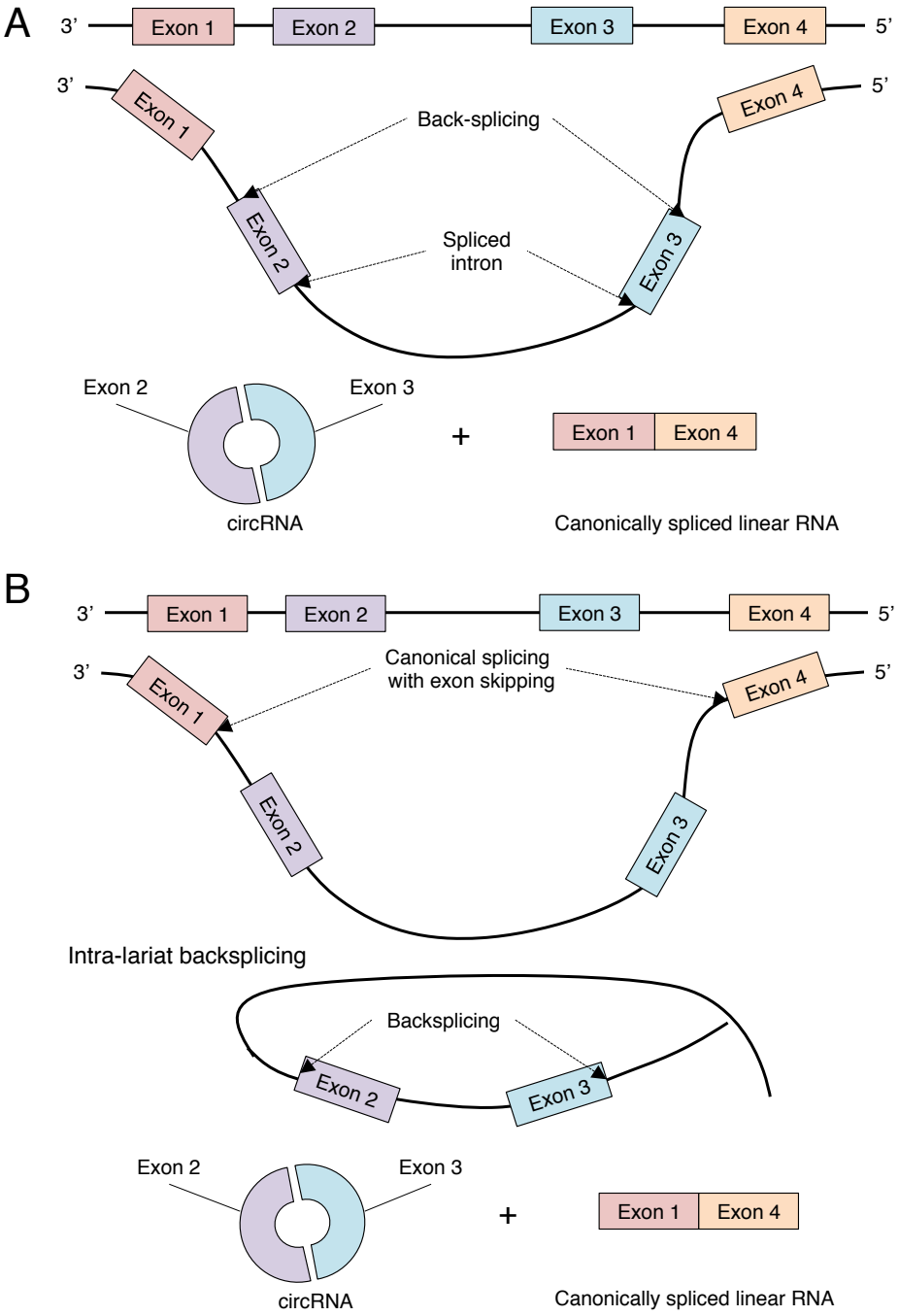

Figure 5.1: **Two circRNA biogenesis pathways thought to occur.** In Figure 5.1A, a model of circRNA biogenesis is depicted. This involves direct back-splicing of the circRNA exons first, then the splicing out of an intervening intron, to produce a circRNA and linear RNA. In Figure 5.1B, the indirect route for circRNA production via exon-skipping is shown. Exons are skipped, and then in the removed lariat, there is back-splicing, resulting in the formation of a circRNA product.

Certain species of circRNA may be involved in protein production, as in addition to the exons for the mRNA gene they are transcribed from, circRNA may also contain an internal ribosome entry site, and therefore may be translated into protein, as in the case of circMbl and the *muscleblind* protein [253, 254]. Other functional roles for circRNA relate to their involvement as miRNA sponges, such as ciRS-7, an endogenously produced circular RNA sponge for miR-7, which has over 70 binding sites on its transcript for this miRNA [29, 247]. This circRNA transcript has shown extensive association with AGO2 in AGO2-CLIP experiments, as well as concurrent association with miR-7, and dramatic changes occur within the transcriptome following its knockdown, suggesting it indeed functions as a sponge for miR-7 within the cell [29, 247]. In addition to ciRS-7, only a handful of other circRNA transcripts have been shown to have strong evidence for sponge effects, such as cir-SRY for miR-138 [247, 255], and cir-ITCH for miR-7, 17, and 214 [256]. Most circRNA are not thought to act as sponges in large part, though many contain small numbers of miRNA binding sites. In addition, circRNA have also been shown to be transcriptional regulators through interactions with the RNA polymerase II complex in the nucleus, resulting in increased transcription of particular genes (e.g. ci-sirt7 and the *SIRT7* gene) [257]. For many circRNA, while at this point in time they are poorly characterised and without functional annotation, this remains an area of increasing research interest.

5.1.2 HSP90 is a molecular chaperone interacting with AGO2

HSP90, initially identified as a highly expressed, evolutionarily conserved protein across species, is well-known in its functional role as a chaperone protein mediating the cellular stress response, particularly to heat [258]. In times of cellular stress, HSP90 undertakes its role and enables differential conformations of cellular proteins, allowing the cell to adapt to the inciting stressor [258, 259]. In characterising the vast family of heat shock proteins, work by Lindquist et al. identified HSP90 as a unique member of this family owing to the diversity of its protein targets, and the crucial roles of these targets in signal transduction, cellular proliferation, survival, and differentiation [259, 260]. As such, HSP90 has been identified as a key mediator of evolution; it can be thought of as a capacitor for evolutionary energy, potentiating evolvability, or the ability to adapt and survive a process, giving rise to natural selection [261, 262, 263, 264].

Recent biochemical work has shown the critical association of HSP90 and AGO2, specifically in mediating the localisation of AGO2 to the sites of aggregated, efficient,

RISC-mediated mRNA repression [265]. This association was identified through live-cell imaging experiments and pharmacological inhibition of HSP90, where it was shown that inhibition of HSP90 uniquely negatively impacted the localisation of AGO2 to stress granules and P-bodies [265]. Furthermore, this HSP90 inhibition has been shown to have a functional effect on miRNA-mediated repression, with Pare et al. showing that miRNA-mediated repression is decreased statistically significantly in cases of HSP90 inhibition [266]. The roles of HSP90 relating to miRNA have also been extended in studies in Drosophila, where HSP90 is thought to enable the more efficient receipt of miRNA duplexes by AGO2 in the RISC formation through conformational changes in AGO2 [267].

5.1.3 Research questions

In this chapter, I ask the question of how the circRNA component of the non-coding transcriptome associates in response to hypoxia, as quantified through the hypoxia gene signature score. Using TCGA breast tumour samples with characterised circRNA and RNA-seq data from MCF-7 cells under normoxic and hypoxic conditions, I ask which circRNA statistically significantly associate with hypoxia, and whether these could affect the change in miRNA biogenesis pathway that has been hypothesised to occur. In addition, because circRNA have not yet been assessed on a large scale from clinical tumour specimens, I use the MCF-7 cell line data to study the circRNA identified through different preparations of RNA samples prior to sequencing, and through different informatics pipelines, to help guide future work in this field.

The remainder of this chapter is structured as follows. In Sections 5.3.1 - 5.3.3, I conduct an analysis of circRNA characterised across the TCGA breast cancer dataset, and identify circRNA that statistically significantly correlate with the hypoxia gene signature score. Next, in Section 5.3.4, I outline the evidence for the involvement of a circRNA antisense to *HSP90AB1* in the hypoxic response among clinical tumour samples. In Sections 5.3.5 - 5.3.8, I study the different circRNA identified through cell line experiments of the MCF-7 cell line in normoxia and hypoxia, and how these differ among various sample preparations and informatics pipelines. Lastly, I show in Section 5.3.9 that the circRNA antisense to *HSP90AB1* is detectable in two of four of the MCF-7 cell line isolates, and of these isolates, the expected response in hypoxia is observed, but owing to small sample size, no conclusions can be drawn definitively.

5.2 Materials and methods

5.2.1 Data sources

5.2.1.1 TCGA circRNA data

To study circRNA expression in human cancers, I analysed the largest publicly-available dataset of circRNA characterised on clinical samples to date. This data series relied upon the raw sequencing information from the samples in the TCGA breast cancer cohort, and mapped the unmapped reads from the RNA-seq data using a circRNA annotation pipeline to identify those arising from putative circRNA [268]. More specifically, the package by Nair et al. for circRNA annotation was called CircSeq, and its approach to identification and annotation of circular RNA is outlined below. This dataset, while being the largest dataset of its kind for clinical tumour samples, suffers from limitations tempering its use and potential for wider application. One is that detection of circRNA from these samples may be technically limited by the protocol used to extract RNA for the TCGA study. The RNA-seq data for the TCGA project was polyA-selected before sequencing, and the unmapped reads reprocessed through the circSeq pipeline by Nair and colleagues was based on this data [268]. Moreover, this dataset was highly sparse, contained a large degree of noise, and suffered from low counts of circRNA across samples (often counts less than 10), and likely only reliably represented the circRNA with the greatest and most consistently detectable expression. It was hypothesised that polyA selection implicitly limited the detectability of circRNA, as these were then depleted by the polyA selective step in sample preparation. To limit the rate of false positive results from this dataset, I used highly stringent initial filtering.

A strict filtering scheme was defined, wherein only those circRNA consistently expressed across at least 20% of breast (invasive ductal carcinoma) tumour samples, were considered. This constraint ensured consistent detectability of circRNA, and reduced the number of total circRNA species from 6104 to 25, thereby removing 99.5% of all species. This set of 25 circRNA was used for all analyses presented.

5.2.2 Experimental methods

5.2.2.1 MCF-7 cell line data

The Buffa and Harris labs previously collaborated to generate genomic data for an ER-positive, progesterone receptor (PR)-positive, HER2-negative, luminal breast cancer cell line, MCF-7, incubated in normoxia (21% oxygen), and hypoxia (1% oxygen).

Cells were incubated under these conditions in the *In vivo2* Hypoxia WorkStation (Ruskinn Technology Ltd., UK). The complete method for the culture conditions for these cells can be found in the methods section of the original publication in which they were studied [269]. Total RNA were isolated from these cells, and sequenced after either ribo-minus preparation (removing all ribosomal RNA), polyA minus preparation (removing all polyadenylated RNA molecules), or polyA selection (selecting for polyadenylated RNA molecules). Paired-end sequencing using the Illumina TruSeq library preparation platform was carried out, and .fastq files produced by next-generation sequencing were subsequently analysed.

5.2.3 Analytical and statistical methods

5.2.3.1 Prognostic analysis

A prognostic analysis to determine the potential clinical significance of each of the 25 circRNA was carried out using linear regression and Cox proportional hazard ratio modelling. This was done using the clinical annotations for the samples from which the circRNA have been quantified, ensuring a fair analysis. The patient demographics of this cohort are listed in Table 5.1, divided by molecular subtype (69% estrogen receptor positive), histologic subtype (74% infiltrating ductal carcinoma), overall stage (58% stage II), nodal stage (46% N0), and tumour stage, based on the tumour size (59% T2). Note that classification for staging purposes is defined in keeping with the conventions outlined in the AJCC 7th edition manual [270].

To detect prognostic ability, the rank-normalised expression levels for the circRNA were used, ranking each of the circRNA expression levels between 0 and 1 for each of the samples, with 0 being the lowest and 1 being the highest expression. A linear modelling scheme was devised to identify those circRNA which were the strongest predictors of prognosis. Rank-normalising these values had several effects. First, it removed the sense of scale, which was preferable in this case, as the sparsity of the data created scales of expression that were different between the various circRNA species. Second, this transformation provided a convenient scaling for each of the predictor variables between 0 and 1. Third, by only considering ranks of the circRNA the modelling scheme emphasised more the ordering of the circRNA than the absolute values or the differences in absolute values between the counts of the circRNA themselves.

Characteristic	No. of patients (%)
Total	710 (100)
Subtype	
ER+	489 (69)
HER2+	155 (22)
TN	66 (9)
Histologic subtype	
Infiltrating ductal	526 (74)
Infiltrating lobular	128 (18)
Mixed	14 (2)
Mucinous	8 (1.1)
Medullary	3 (0.4)
Other	31 (4.4)
Stage	
I	115 (16)
II	411 (58)
III	167 (24)
IV	17 (2.4)

Characteristic	No. of patients (%)
Nodal stage	
N0	329 (46)
N1	248 (35)
N2	81 (11)
N3	44 (6.2)
Nx	8 (1.1)
Tumour stage	
T1	165 (23)
T2	420 (59)
T3	103 (15)
T4	22 (3.1)

Table 5.1: **Patient characteristics as considered in prognostic analysis.** Table shows the breakdown of patient characteristics considered. ER+ refers to estrogen receptor positive tumours, HER2+ to *HER2* amplified tumours, and TN to triple negative tumours. Staging is reported by the conventions outlined in the AJCC 7th Ed.

5.2.3.2 Linear modelling

A linear modelling approach was used to examine the relationship of circRNA expression to hypoxia gene signature score for the TCGA breast tumour samples considered. Because the circRNA were largely co-correlated, to identify their associations with the hypoxia gene signature score, a linear modelling approach with a L1/L2 penalisation term was employed. In this case, due to the relatively low number of predictors present, a preliminary univariate feature selection scheme was not necessary. The model itself relied upon elastic net regression in order to maximise the benefits of using a L1 penalty to shrink coefficients to zero where possible, and prevent overfitting, and an L2 penalty to reduce the effects of co-correlated circRNA. Moreover, this approach was advantageous in helping to prevent overfitting in this scenario, as circRNA expression was sparse, and the risk of high co-correlation and overfitting was high. For each of the circRNA predictors in the linear model, their expression was scaled by the z transform first to have zero mean and unit variance. The same scaling was done for the hypoxia gene signature score for each of the corresponding samples, where gene signature score was defined as the median expression of all sig-

nature genes. Using combined L1/L2 penalised linear regression, a linear model was fit using each of the 25 conserved circRNAs as predictors, and then 10-fold cross-validation was used in identifying the optimal model fitting the hypoxia score to this linear model.

After coefficients were obtained, tests of statistical significance were not performed, per convention for penalised linear regression models. Standard errors, from which confidence intervals are based, are not meaningful in this context. The effect of penalty terms in the linear model reduces the possible variance of the coefficients of the linear model in a manner that is not possible to predict. Thus, without an understanding of the true possible variance of the model coefficients, an understanding of their underlying distribution cannot be obtained, and therefore, confidence intervals cannot reliably be determined. Further discussion on this convention can be found in the documentation of the penalized R package that was used in the implementation of these linear models [130].

5.2.4 Computational methods

5.2.4.1 Vienna RNAfold

The Vienna RNAFold web server is a portal by which DNA or RNA sequences can be examined for their predicted secondary structures based on a number of biochemical factors, including complementary base pairing, hydrostatic interactions, polar interactions, the effects of temperature, in order to minimise total free energy of the structure. In this way, it is designed to predict a secondary structure conformation that may represent what is occurring within the cell, and it does so through a dynamic programming algorithm described in detail by Zuker et al., wherein different loops and external bases in the RNA molecule are optimised separately, and later combined for a final prediction [271].

5.2.4.2 circRNA identification pipelines

In order to study the detection of circRNA among these samples in the different preparations, I implemented two distinct circRNA identification pipelines, as summarised by the approach in Figure 5.2. For both pipelines, adapter trimming using Trim Galore (Babraham bioinformatics [272]) was performed first, and quality control checks on the fastq files using FastQC were performed next. Throughout this process, it was ensured that the RNA sequencing data passed quality-control standards [273].

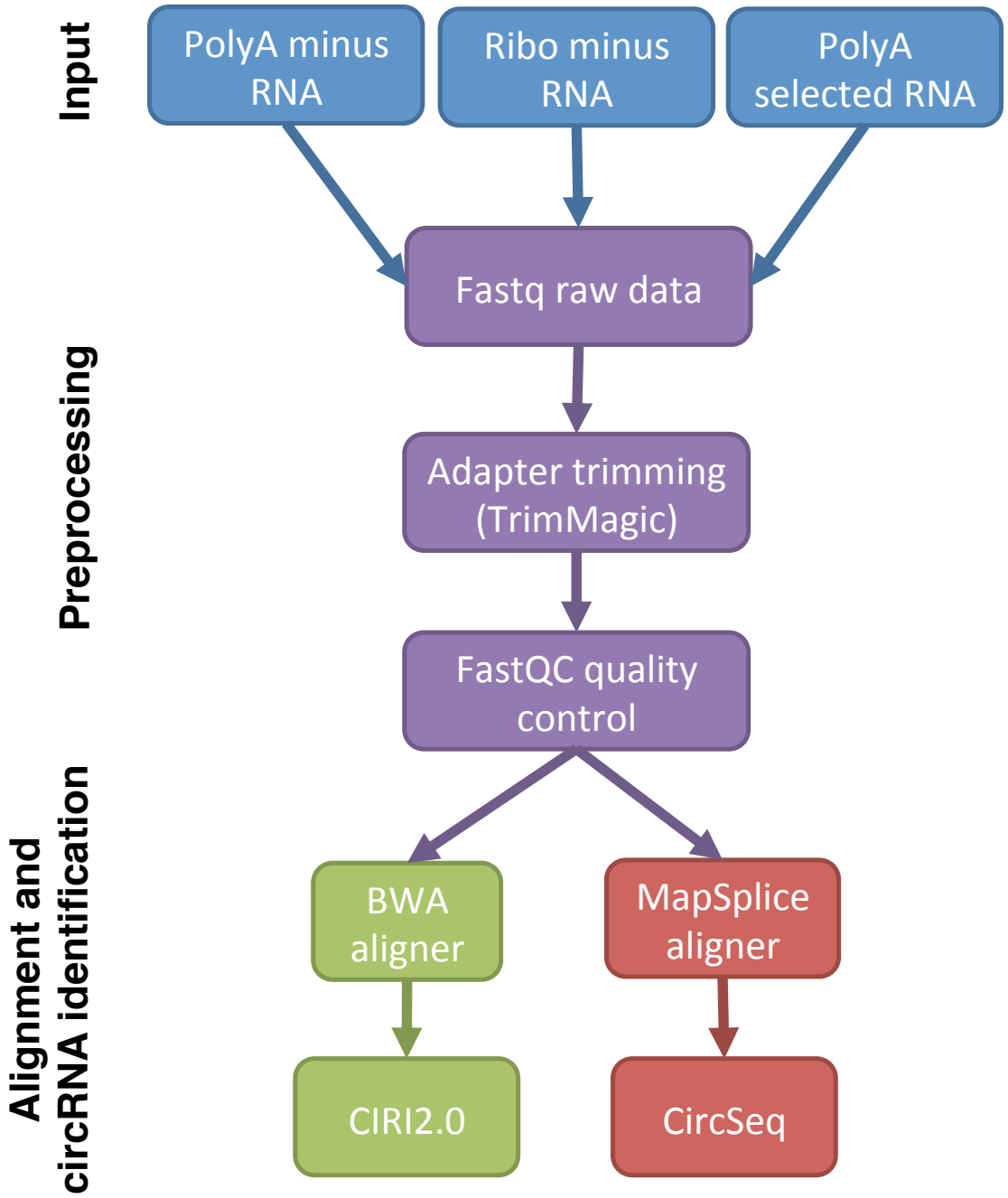

Figure 5.2: **Two distinct circRNA identification algorithms were employed in analysis of the MCF-7 cell line data.** Flow diagram depicting the steps and tools implemented at each stage of computational pipeline for circRNA identification, depicting the workflow for the circSeq and CIRI pipelines.

The first pipeline implemented CIRI2.0, used through the integrated miARma-seq package [193]. In brief, this pipeline characterises circRNA from unmapped reads spanning junctional sites that are potentially generated by back-splicing. It implements a maximum likelihood estimation of the likelihood that a given read comes from either a forward splice junction (i.e. alternatively spliced mRNA) or a back-spliced junction (i.e. from a circRNA). This likelihood is computed by taking the unmapped read, and then breaking it down into smaller seed components, and determining the likelihood that each of the small seed portions of the read come from either the forward or back-spliced junctions. Iterating this process, and then determining the overall likelihood from either scenario, enables CIRI2.0 to determine whether it is more likely that a given unmapped read has come from a back-spliced junction or a forward splice junction, and thereby whether the read is possibly from a circRNA [274].

The miARma-seq pipeline implements FastQC as an initial check pre-alignment for quality control of the fastq files, and subsequently uses the BWA aligner for alignment to the reference genome [55]. For alignment, in order to maintain consistency with the results of the alignment done with the TCGA dataset by Nair et al. in [268], the hg19/GRCh37.75 reference genome was used. CIRI2.0 was chosen for its high sensitivity in detecting *de novo* circRNA, as reported in previous studies comparing the various circRNA identification tools [274]. The generated counts were then normalised to reads per million (RPM) values by dividing raw read counts for each sample by a scaling factor equal to the total number of reads in the sample divided by 10^6.

The second circRNA identification pipeline implemented was circSeq, the same pipeline used by Nair et al. for the mapping of the unmapped reads from TCGA alignment files to putative circRNA backspliced junction sites [268]. CircSeq has been designed to first segment unmapped reads into smaller (20 base pair) anchors, which are then re-mapped to the genome. Those that re-map in a 3' - 5' direction are taken as potentially back-spliced reads, and are used as anchors for evidence for potential circRNA. These are checked for whether the sequences contain possible splice donor and acceptor sequences (AG and GT), and if so, this is considered a circRNA candidate. In this case, the total number of anchor-based reads for each of the circRNA candidates is calculated, and if it surpasses a user-defined threshold expression filter, a size filter, and a validation filter, then the species is considered a true circRNA. Any putative circRNA expressed at a count of less than 5 were excluded. The size filter discarded any putative circRNA that were less than 6 bp in

length. The validation filter excluded any circRNA from repetitive regions of the genome, and ensured that predicted start and end sites were unique. To maintain consistency with Nair et al., prior to using circSeq, the MapSplice aligner was used with the recommended circRNA detection settings (minimum fusion distance 200bp, GRCh37.75 gtf file), to obtain the unmapped reads in the fastq file, for the reference hg19/GRCh37.75 genome [268, 275]. Next, the same base settings for circRNA alignment as Nair and colleagues were used on the unmapped reads to obtain the counts of circRNA in each sample of the MCF-7 data. Raw counts were RPM normalised using the same scaling factor method as above.

Once the RPM normalisation was carried out, counts were log-transformed using the transformation $\log_2(x+1)$, for a RPM normalised count x. These transformed, normalised counts were used for all further downstream analyses of the circRNA.

5.3 Results

5.3.1 Expression of circRNA across breast cancer samples

After the initial cleaning of the circRNA dataset from Nair et al., the characteristics of the expression of the circRNA under consideration were examined. A heatmap for the expression of these conserved circRNA is shown in Figure 5.3a, noting that circRNA with fewer than 5 reads in a given sample were also removed as a part of the processing algorithm implemented by Nair et al. [268].

Next, the genomic origin of these circRNA was analysed, as shown in Figure 5.3b, where both the proportion of the circRNA detected, and the proportion of expressed species (weighting more highly for more highly expressed circRNA), are shown. The filtered circRNA (most consistently expressed 25 circRNA) were compared to the unfiltered distributions, to analyse for potentially biased results. With respect to the distribution of the circRNA detected, a bias was observed towards circRNA originating from chromosomes 6 and 14 that is not present in the overall, unfiltered dataset. This may have been due to the high expression of circRNA from these chromosomes, relative to circRNA from other chromosomes, as evidenced by the distribution of reads by chromosome on the lower chart of Figure 5.3b, which also shows a bias towards chromosomes 6 and 14, even in the unfiltered case. Another possibility is that highly expressed circRNA in these samples arose primarily from these chromosomes, or that these sequences tended to be more stable during sequencing, though these remain hypotheses.

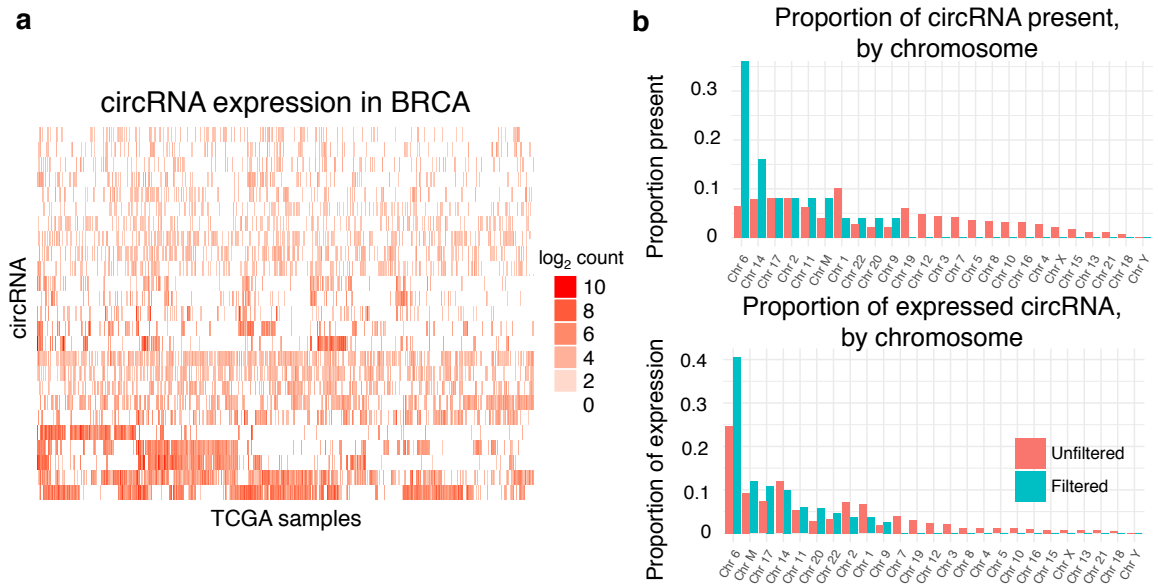

Figure 5.3: **Expression profile of circRNA consistently expressed in TCGA breast tumours.** A) Expression profile showing log count of the circRNA identified through the analysis by Nair et al. for circRNA in the TCGA breast cancer dataset. B) Bar graphs depicting the composition of all circRNA (unfiltered, orange) and the consistently expressed circRNA (filtered, teal) by chromosome of origin (upper), and weighted by expression of each circRNA (lower), by chromosome of origin. Note that Chr M refers to the mitochondrial genome.

For 8 of these 25 circRNA, the sequence length was within the range that the secondary structure could be predicted by the RNAFold algorithm from ViennaRNA [276]. Using the RNAFold web server, the secondary structures maximising the free energy of the RNA molecule, at 37 degrees, with default settings, was computed. The molecule was assumed to be circular in this computation. This showed that among even the 8 circRNA for which this computation could be done, a diversity of structures and motifs is apparent, as depicted in Figure 5.4. Lastly, also from this figure, the effects of polyA selection are apparent, as the secondary structures of these circRNA revealed loops of A-rich sequences that may have been the loci captured by the selection step during sample preparation.

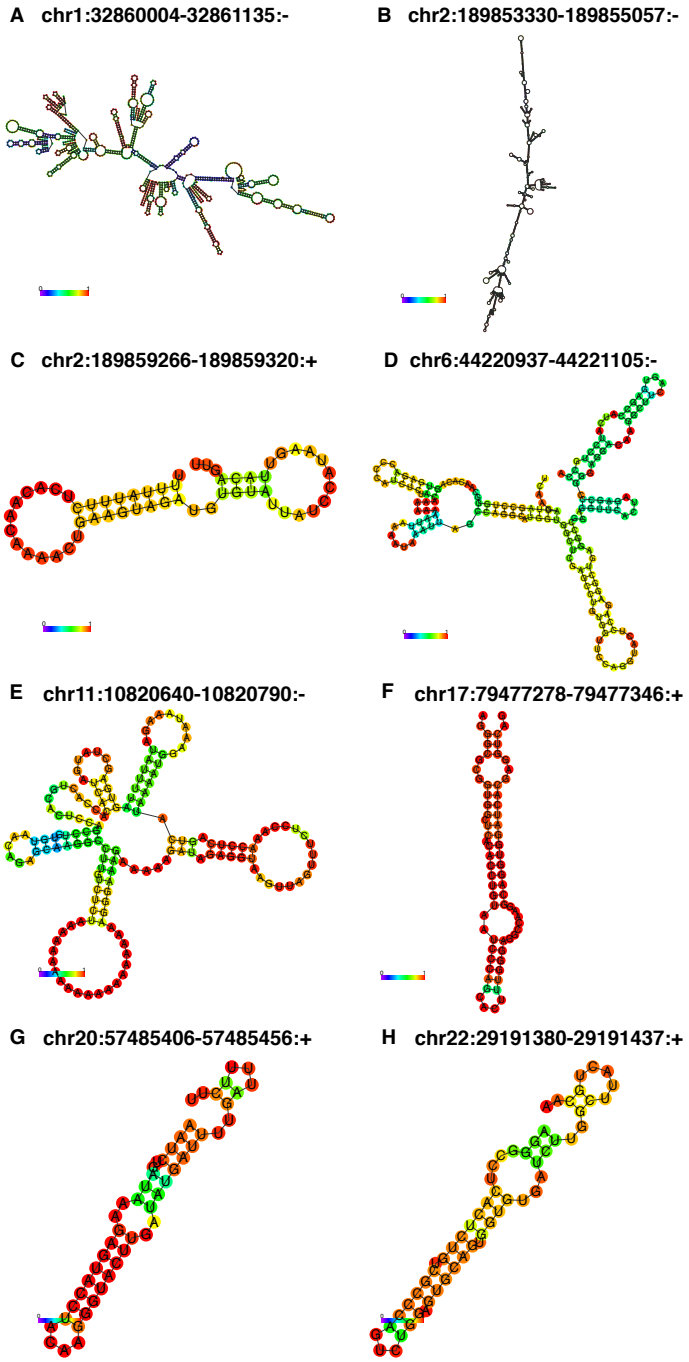

Figure 5.4: **circRNA identified show a diversity of predicted secondary structures.** Secondary structures were predicted to maximise free-energy as computed by the Vienna RNAFold software. Nucleotides are coloured by probability of binding. Among many of these molecules, an A-rich region potentially serving as the basis of its selection in polyA selected sequencing can be seen.

5.3.1.1 Genes antisense to circRNA show highly correlated expression

To further analyse the genomic origins of these circRNA, the overlapping genes from their loci are listed in Table 5.2, where possible. Approximately two-thirds of the consistently expressed circRNA were found to overlap exons, with the remainder being circular intronic species. One circRNA, chr14:106144679-106207906:- was found to overlap a pseudogene, *ELK2AP* on the sense strand.

The rank (Spearman) correlation coefficient between expression of the overlapping genes and their corresponding circRNA was calculated. This is graphically depicted for an example circRNA in Figure 5.5. In some cases, these circRNA which overlap with a gene, either in sense or antisense, were statistically significantly positively correlated, and statistically significant negative correlation was only observed in the case of one circRNA, chr17:47419917-48261732:+ for the genes *PHB*, *DLX3*, *SPOP*, *SLC35B1*, *ZNF652*, *FAM117A*, *PPP1R9B*, *TAC4*, *LOC284080*, and *HILS1* ($\rho = -0.17$, $p < 10^{-4}$ for all genes listed). The strongest positive correlations were observed for the antisense genes overlapping a circRNA, suggesting a possible role for these circRNA in the stabilisation or coexpression of these genes. The potential role of circRNA stabilising antisense transcripts to the genes from which they are transcribed was experimentally confirmed in a recent report of the CDR1as circRNA found to stabilise the *CDR1* gene through a strand-specific quantisation [277]. Additionally, for another form of non-coding RNA, an analogous finding has also been reported; the antisense *BACE1as* transcript has been shown to stabilise the *BACE1* gene by way of an RNA duplex mediated by complementary base pair binding [278].

Table 5.2: **circRNA show highly statistically significant correlations with overlapping genes on sense and antisense strands.** Spearman correlation computed for circRNA and the linear transcripts overlapping on the corresponding sense and antisense strands. Only circRNA with at least one overlapping gene are listed in this table.

circRNA	Sense genes	Antisense genes
chr11:10820640-10820790:-	*EIF4G2* $\rho = 0.24$, $p < 10^{-9}$	
chr1:32860004-32861135:-	*BSDC1* $\rho = 0.31$, $p < 10^{-13}$	
chr14:67708094-69255647:+	*ARG2* *EIF2S1* *RAD51B* *PLEKHH1* *MPP5* *RDH12* $\rho = 0.13$, $p = 0.001$	*ZFP36L1* *PIGH* *VTI1B* *ZFYVE26* *PLEK2* *RDH11* *ATP6V1D* *TMEM229B* $\rho = 0.35$, $p < 10^{-18}$
chr2:189859266-189859320:+	*COL3A1* $\rho = 0.35$, $p < 10^{-18}$	
chr6:32489681-32522748:-	*HLA-DRB5* $\rho = 0.12$, $p = 0.003$	
chr6:32497901-32525963:-	*HLA-DRB5* $\rho = 0.08$, $p = 0.06$	
chr6:44220937-44221105:-		*HSP90AB1* $\rho = 0.43$, $p < 10^{-28}$
chr14:106134679-106207906:-	*ELK2AP* (NA)	
chr17:47419917-48261732:+	*ITGA3* *DLX4* *NGFR* *PDK2* *SGCA* *KAT7* *NXPH3* *FLJ45513* $\rho = 0.16$, $p = 0.0001$	*COL1A1* $\rho = 0.48$, $p < 10^{-35}$ *SAMD14* $\rho = 0.18$, $p < 10^{-5}$ *PHB* *DLX3* *SPOP* *SLC35B1* *ZNF652* *FAM117A* *PPP1R9B* *TAC4* *LOC284080* *HILS1* $\rho = -0.17$, $p < 10^{-4}$
chr22:29191380-29191437:+		*XBP1* $\rho = 0.62$, $p < 10^{-65}$
chr20:57485406-57485456:+	*GNAS* $\rho = 0.005$, $p = 0.9$	
chr2:189853330-189855057:-		COL3A1 $\rho = 0.54$, $p < 10^{-45}$
chr6:32522516-32549383:+		*HLA-DRB5* $\rho = 0.004$, $p = 0.91$
chr17:79477278-79477346:+		*ACTG1* $\rho = 0.33$, $p < 10^{-16}$

5.3.1.2 circRNA associate primarily positively in expression with miRNA

For each of the 25 circRNA identified as consistently expressed across breast cancers, their associated miRNA expression was also examined. For this, the most statistically significantly positively and negatively correlated miRNA with each circRNA were identified using the Spearman correlation coefficient. These results are summarised in Table 5.3.

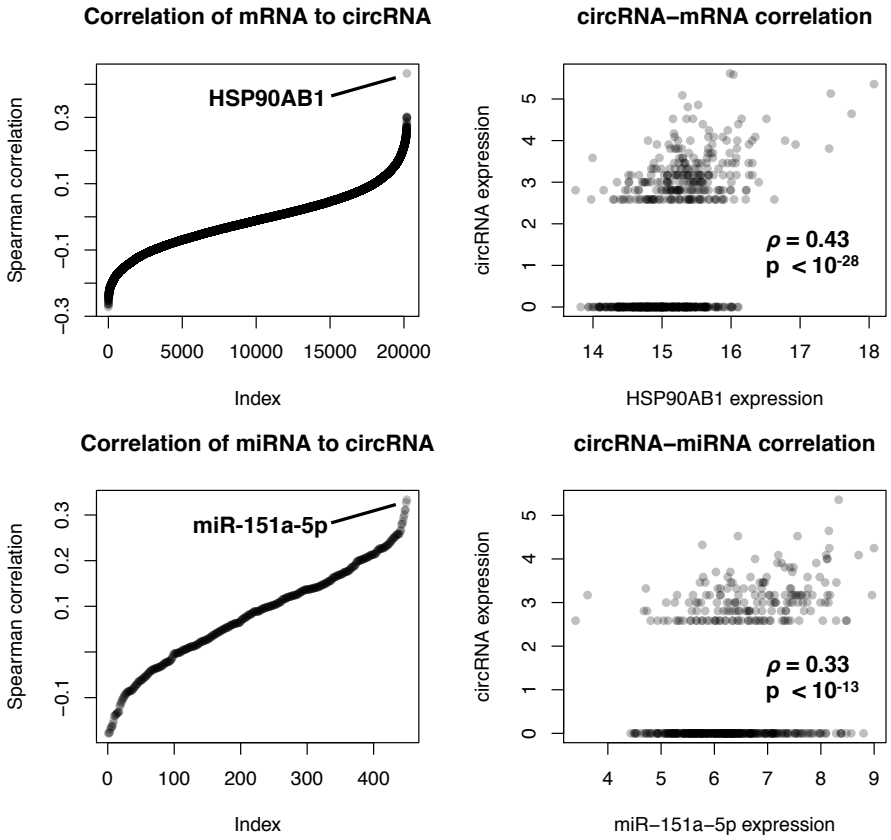

Figure 5.5: **Sample of correlation plots depicting circRNA-mRNA and circRNA-miRNA correlations.** Left: Plots depicting the range of Spearman correlation coefficients for circRNA-mRNA correlations (top) and circRNA-miRNA correlations (bottom). Right: Example plots showing the circRNA and mRNA levels for the most correlated species (top), and likewise for the circRNA and miRNA (bottom). These plots are shown for the circRNA chr6:44220937-44221105:-, antisense to *HSP90AB1*.

Table 5.3: **circRNA may show statistically significant positive or negative associations with miRNA.** Spearman correlation computed for circRNA and all miRNA, with up to the top 5 miRNA showing either positive or negative statistically significant correlations listed. Only circRNA with at least one statistically significantly correlated miRNA are listed in this table.

circRNA	Positively associated miRNA	Negatively associated miRNA
chr11:10820640-10820790:-	628-5p $\rho = 0.30$, $p < 10^{-10}$ 26a-1-3p $\rho = 0.27$, $p < 10^{-8}$ 3605-3p $\rho = 0.24$, $p < 10^{-6}$ 148b-3p $\rho = 0.22$, $p < 10^{-6}$ 186-5p $\rho = 0.22$, $p < 10^{-6}$	
chr1:32860004-32861135:-	29b-2-5p $\rho = 0.22$, $p < 10^{-6}$	106b-5p $\rho = -0.25$, $p < 10^{-7}$ 214-3p $\rho = -0.23$, $p < 10^{-6}$ 18a-5p $\rho = -0.21$, $p < 10^{-5}$ 27b-3p $\rho = -0.20$, $p < 10^{-5}$ 376c-3p $\rho = -0.20$, $p < 10^{-4}$
chr2:189859266-189859320:+	493-5p $\rho = 0.23$, $p < 10^{-6}$ 199a-5p $\rho = 0.22$, $p < 10^{-5}$ 379-5p $\rho = 0.21$, $p < 10^{-5}$	106b-5p $\rho = -0.25$, $p < 10^{-7}$ 32-5p $\rho = -0.24$, $p < 10^{-7}$ 141-3p $\rho = -0.24$, $p < 10^{-7}$ 590-5p $\rho = -0.23$, $p < 10^{-6}$ 30e-5p $\rho = -0.23$, $p < 10^{-6}$
chr6:32489681-32522748:-	150-5p $\rho = 0.23$, $p < 10^{-6}$ 142-5p $\rho = 0.22$, $p < 10^{-5}$ 142-3p $\rho = 0.21$, $p < 10^{-5}$ 155-5p $\rho = 0.20$, $p < 10^{-5}$ 150-3p $\rho = 0.21$, $p < 10^{-5}$	
chr6:44220937-44221105:-	151a-5p $\rho = 0.33$, $p < 10^{-13}$ 937-3p $\rho = 0.33$, $p < 10^{-12}$ 301a-3p $\rho = 0.31$, $p < 10^{-11}$ 940 $\rho = 0.30$, $p < 10^{-10}$ 877-5p $\rho = 0.31$, $p < 10^{-10}$	
chr9:19380187-88990615:-	101-3p $\rho = 0.24$, $p < 10^{-7}$ 455-5p $\rho = 0.22$, $p < 10^{-6}$ 19b-1-5p $\rho = 0.21$, $p < 10^{-5}$ 20a-5p $\rho = 0.21$, $p < 10^{-5}$ 19b-3p $\rho = 0.21$, $p < 10^{-5}$	375 $\rho = -0.19$, $p < 10^{-4}$
chr14:106134679-106207906:-	155-5p $\rho = 0.30$, $p < 10^{-11}$ 150-5p $\rho = 0.27$, $p < 10^{-9}$ 146a-5p $\rho = 0.27$, $p < 10^{-9}$ 150-3p $\rho = 0.25$, $p < 10^{-7}$ 142-5p $\rho = 0.23$, $p < 10^{-6}$	

Table 5.3: **circRNA may show statistically significant positive or negative associations with miRNA.** Spearman correlation computed for circRNA and all miRNA, with up to the top 5 miRNA showing either positive or negative statistically significant correlations listed. Only circRNA with at least one statistically significantly correlated miRNA are listed in this table.

circRNA	Positively associated miRNA	Negatively associated miRNA
chr17:47419917-48261732:+	134-5p $\rho = 0.42$, $p < 10^{-22}$ 539-5p $\rho = 0.37$, $p < 10^{-16}$ 493-5p $\rho = 0.37$, $p < 10^{-16}$ 409-3p $\rho = 0.34$, $p < 10^{-14}$ 337-3p $\rho = 0.34$, $p < 10^{-14}$	101-3p $\rho = -0.35$, $p < 10^{-15}$ 17-3p $\rho = -0.35$, $p < 10^{-15}$ 20a-5p $\rho = -0.32$, $p < 10^{-12}$ 106b-5p $\rho = -0.31$, $p < 10^{-12}$ 29a-3p $\rho = -0.30$, $p < 10^{-11}$
chr22:29191380-29191437:+	190b $\rho = 0.32$, $p < 10^{-12}$ 29b-2-5p $\rho = 0.31$, $p < 10^{-12}$ 29c-5p $\rho = 0.27$, $p < 10^{-9}$ 375 $\rho = 0.27$, $p < 10^{-9}$ 153-3p $\rho = 0.23$, $p < 10^{-6}$	17-3p $\rho = -0.37$, $p < 10^{-17}$ 18a-5p $\rho = -0.37$, $p < 10^{-17}$ 378a-5p $\rho = -0.37$, $p < 10^{-17}$ 378a-3p $\rho = -0.35$, $p < 10^{-15}$ 660-5p $\rho = -0.35$, $p < 10^{-14}$
chr6:31237776-31320721:-		361-5p $\rho = -0.20$, $p < 10^{-5}$
chr2:189853330-189855057:-	199a-3p $\rho = 0.37$, $p < 10^{-16}$ 199b-3p $\rho = 0.37$, $p < 10^{-16}$ 409-3p $\rho = 0.37$, $p < 10^{-16}$ 382-5p $\rho = 0.37$, $p < 10^{-16}$ 134-5p $\rho = 0.37$, $p < 10^{-16}$	200b-3p $\rho = -0.26$, $p < 10^{-8}$ 106b-5p $\rho = -0.26$, $p < 10^{-8}$ 200c-3p $\rho = -0.25$, $p < 10^{-7}$ 429 $\rho = -0.24$, $p < 10^{-7}$ 200a-3p $\rho = -0.23$, $p < 10^{-6}$
chr6:32522516-32549383:+	150-3p $\rho = 0.21$, $p < 10^{-5}$	
chr6:33053555-33095875:+	146a-5p $\rho = 0.21$, $p < 10^{-5}$ 378c ($\rho = 0.20$, $p < 10^{-5}$	
chrM:10090-10164:-	195-5p $\rho = 0.26$, $p < 10^{-8}$ 26b-5p $\rho = 0.23$, $p < 10^{-6}$ 365a-3p $\rho = 0.23$, $p < 10^{-6}$ 423-5p $\rho = 0.22$, $p < 10^{-6}$ 26a-5p $\rho = 0.21$, $p < 10^{-5}$	16-1-3p $\rho = -0.23$, $p < 10^{-6}$
chr14:106208067-106235831:-	155-5p $\rho = 0.30$, $p < 10^{-10}$ 142-5p $\rho = 0.27$, $p < 10^{-8}$ 146a-5p $\rho = 0.24$, $p < 10^{-7}$ 150-5p $\rho = 0.23$, $p < 10^{-6}$ 150-3p $\rho = 0.21$, $p < 10^{-5}$	
chr14:106209355-106237690:+	155-5p $\rho = 0.44$, $p < 10^{-24}$ 146a-5p $\rho = 0.38$, $p < 10^{-17}$ 142-5p $\rho = 0.37$, $p < 10^{-16}$ 150-5p $\rho = 0.36$, $p < 10^{-16}$ 150-3p $\rho = 0.35$, $p < 10^{-13}$	190b $\rho = -0.19$, $p < 10^{-4}$

circRNA	miRNA binding sites (number)
chr1:32860004-32861135:-	4266 (2), 4668-5p (1), 4534 (2), 23 others
chr2:189853330-189855057:-	875-3p (3), 4306 (1), 194-3p (1), 1299 (3), 4795 (4), 40 others
chr6:44220937-44221105:-	1304-3p (2)
chr11:10820640-10820790:-	504-3p (1), 4531 (1), 4439 (1), 2 others
chr14:106208067-106235831:-	5011-5p (1)

Table 5.4: **circRNA containing miRNA binding sites.** Table lists the circRNA and potential miRNA for which they have target sites, and their number. This was only done for circRNA whose sequences were within 100-30,000 nucleotides long, as per the requirements of the miRDB database. In the case of multiple miRNA binding sites, the three with the highest target likelihood score are listed, as well as those occurring at least three times.

From this table, it may be observed that for most circRNA, there are a number of miRNA for which their expression correlates statistically significantly. Indeed, Table 5.3 shows that, of the 25 circRNA considered, there are 10 species that correlated primarily positively with miRNA, 2 that correlated primarily negatively with miRNA, 4 that correlated statistically significantly both positively and negatively with miRNA, and 9 that did not correlate statistically significantly with miRNA expression for any miRNA. Thus, within the limitations of this analysis, as might be expected from previous reports, circRNA do not appear to function as miRNA sponges on a large scale, as in this case, more negatively correlated miRNA would be expected; rather, based on these results, miRNA sponges may only be limited to specific species of circRNA. The statistically significant positive associations observed between most circRNA and miRNA may have been due to either co-transcription, stabilisation, interactions through a common repressed target, or more complex ceRNA interactions, though experimental validation is required to confirm these hypotheses.

Moreover, for the circRNA whose sequence length was within the range 100-30,000 nucleotides, the miRDB database and target prediction algorithm were used to check for miRNA target sites [139, 279]. The results of this are summarised in Table 5.4. For those circRNA that were possible to check, given the size constraints, these did harbour many miRNA binding sites, in some cases, up to four for the same miRNA. However, these remain theoretical, sequence-based predictions, and many of the miRNA predicted to be targeted appear to be either poorly expressed or very rare.

5.3.2 Prognostic analysis of circRNA expression

To determine the clinical relevance of the 25 consistently expressed circRNA, an analysis for overall survival as compared to their expression was carried out. As

described in the Methods section above, this was done through three analyses: in the first analysis, each circRNA was considered as a sole predictor in a univariate Cox proportional hazards model for the overall survival among the TCGA cohort. In the second analysis, these hazard ratios were adjusted for other crucial clinical variables; namely, molecular subtype, histologic subtype, overall clinical stage, nodal stage, and tumour stage. In the third part of this analysis, a general linear model involving all of the 25 circRNA was considered in a Cox proportional hazards model, but also whilst adjusting for the aforementioned clinical variables, to generate a predictive classifier.

Figure 5.6 summarises the results obtained from the single circRNA analyses, and depicts the hazards ratios as determined for each of the circRNA individually, and individually when adjusted for molecular and histologic subtypes, and clinical, tumour, and nodal stages. These results reveal that there are a number of circRNA of those considered that indeed do show prognostic significance, even after adjustment for relevant clinical parameters. More specifically, there are seven circRNA that show significance in the single circRNA analysis of survival, after adjustment for the clinical variables considered, suggesting that they may be clinically relevant. The circRNA which were found to have expression associated with improved prognosis are as follows, listed with associated Cox proportional hazards ratios, 95% confidence interval, and p value, as follows: chr14:106209355-106237690:+ 0.15 $(0.03 - 0.73, p = 0.02)$, chr6:33053555-33095875:+ 0.16 $(0.03 - 0.73, p = 0.02)$, and chr14:106134679-106207906:- 0.19 $(0.04 - 0.91, p = 0.04)$. Likewise, circRNA which were found to have expression associated with poorer prognosis are: chr11:10820640-10820790:- 6.84 $(2.55 - 18.35, p < 0.01)$, chr14:67708094-69255647:+ 3.88 $(1.19 - 12.66, p = 0.02)$, chr6:44220937-44221105:- 3.11 $(1.1 - 8.8, p = 0.03)$, and a purported mitochondrial circRNA, chrM:2821-2880:- 2.97 $(1.12 - 7.88, p = 0.03)$.

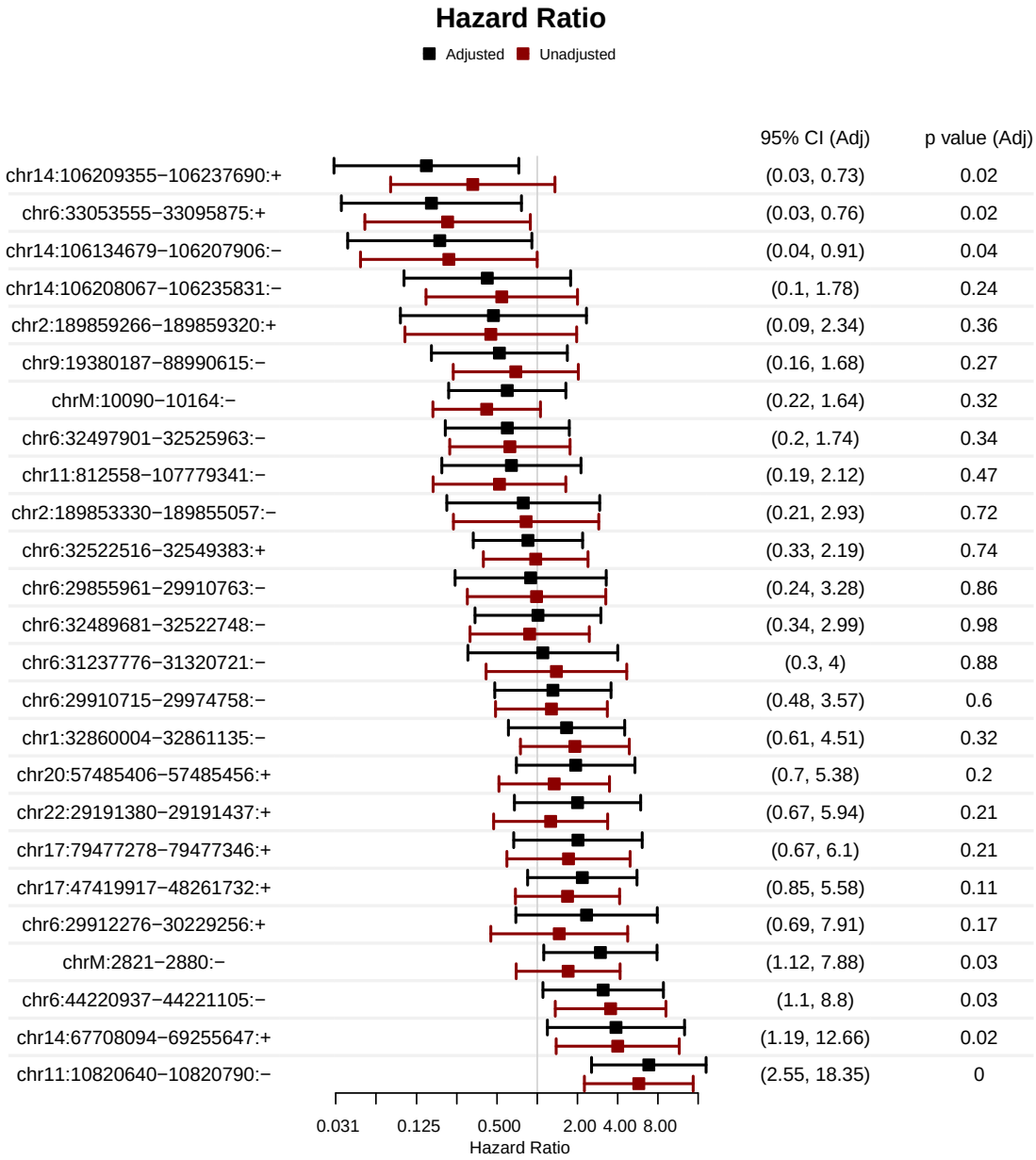

Figure 5.6: **Forest plot of hazard ratios for circRNA expression and prognosis in an individual circRNA Cox proportional hazards model for overall survival.** 95% confidence intervals are shown. Intervals shown in black have been adjusted for molecular and histologic subtype, as well as tumour, clinical, and nodal stage. circRNA expression was rank normalised before model fitting. Intervals in maroon were not adjusted for any clinical covariates, and represent the univariate association between overall survival and circRNA expression in a Cox proportional hazards model. The numeric confidence intervals and p values shown are for the adjusted (black) confidence intervals only.

Following this, a linear model involving all of the 25 circRNA was considered as a predictor for overall survival, again in a Cox proportional hazards model. In considering this model, the linear combination of rank normalised circRNA expression was adjusted for key clinical variables, as in the previous analysis. Namely, the combined linear model was adjusted for the molecular and histologic subtype, as well as the tumour, clinical, and nodal subtypes. The results of this analysis are shown in Figure 5.7, and reveal that, for the combined predictor, when adjusted for all other clinical variables, three circRNA showed statistically significant hazard ratios, suggesting their strength above others as drivers in this linear model. More specifically, chr6:33053555-33095875:+ showed association with better prognosis, with hazard ratio 0.11 $(0.02-0.74, p = 0.02)$, and chr11:10820640-10820790:- 5.72 $(1.79-18.3, p < 0.01)$ and chrM:2821-2880:- 5.37 $(1.55-18.55, p = 0.01)$ showed association with poor prognosis. Among the clinical covariates, the HER2+ molecular subtype showed statistically significant association with poor prognosis with hazard ratio 4.23 $(1.99-8.96, p < 0.01)$, and the infiltrating lobular histologic subtype showed statistically significant association with poor prognosis as well, with hazard ratio 4.50 $(2.13-9.5, p < 0.01)$, but no other clinical predictors were found to be statistically significant as covariates with the combined linear model.

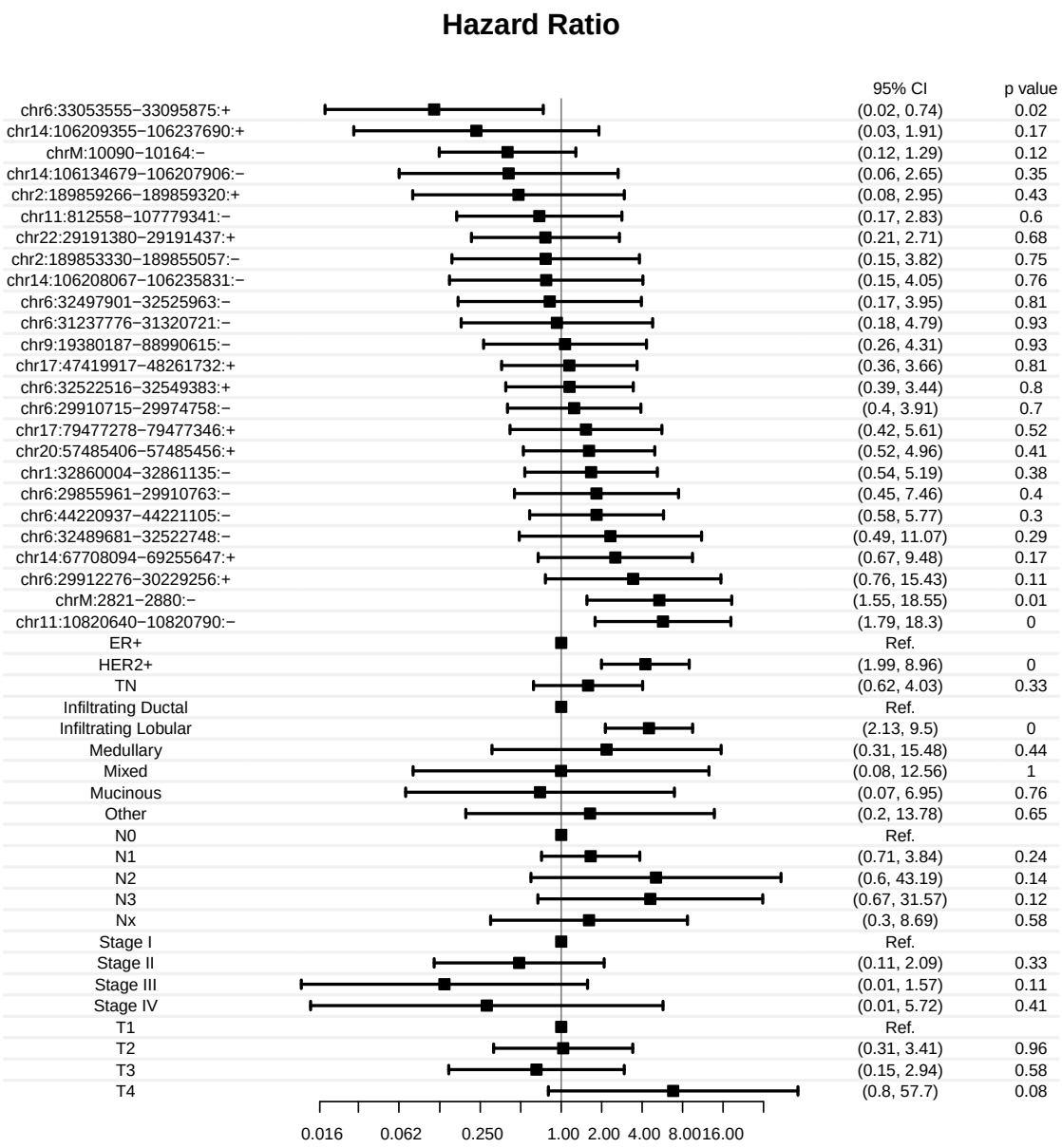

Figure 5.7: **Forest plot of hazard ratios for circRNA expression and prognosis as combined predictor in a Cox proportional hazards model.** Confidence intervals shown are for the hazard ratios as fit to each covariate in the general linear model involving circRNA expression of the 25 circRNA under consideration, and adjusted for histologic and molecular subtype, tumour, clinical, and nodal stage. circRNA expression was rank normalised before model fitting.

5.3.3 circRNA showing association with hypoxia gene signature score

Next, the association of circRNA with hypoxia, as based upon the gene signature score published by Buffa et al., was determined. As described in the Methods, a penalised linear model was fit, using the z transformed expression of the circRNA as predictors, and the z transformed hypoxia gene signature score as the response variable, to determine the association of the circRNA to the hypoxia score. This model was fit using elastic net linear regression, and with a combined univariate-multivariate approach, as described above.

After employing this approach to model fitting, of the 25 circRNA in the model, 16 had non-zero coefficients, and these are listed in Table 5.5, below, ranked by the magnitude of their coefficient in the linear model.

The circRNA that showed the strongest negative association with hypoxia score were those antisense and sense to *COL3A1*, antisense to *XBP1*, and sense to *BDSC1*. Further, the circRNA with the strongest positive associations with hypoxia were those antisense to *HSP90AB1* and *ACTG1*, and sense to *HLA-DRB5* and *GNAS*. The genes *XBP1* and *HSP90AB1* are both genes involved in different cellular stress responses; namely, the endoplasmic reticulum stress response in the case of *XBP1* [280], and the heat shock response in the case of *HSP90AB1* [281]. This suggested that the hypoxia gene signature, when mapped to the level of circRNA, may associate with circRNA-overlapping genes potentially associated with the cellular stress response.

5.3.4 A circRNA antisense to *HSP90AB1* correlates in expression with hypoxia gene signature score and *AGO2* expression

From the previous analyses, a number of circRNA that were consistently detected across breast tumour samples, as well as those that correlated with tumour hypoxia and were predictive of prognosis, were identified. In particular, the circRNA at the locus chr6:44220937-44221105:-, lying antisense to the *HSP90AB1* gene, was expressed at detectable levels in 39.1% of TCGA breast tumour samples, associated positively with hypoxia score, and was predictive of poor prognosis even after adjustment for related clinical variables (stage, tumour stage, nodal stage, histologic and molecular subtype), in a Cox proportional hazards model. Next, the relationship between the expression of each of these circRNA with the expression of *AGO2* mRNA was examined. To do this, the Spearman correlations were computed for each of the 25

circRNA	Coefficient	Overlapping genes
Chr22:29191380-29191437:+	-76.9	*XBP1-as*
Chr2:189859266-189859320:+	-22.0	*COL3A1*
Chr1:32860004-32861135:-	-15.7	*BDSC1*
Chr2:189853330-189855057:-	-11.4	*COL3A1-as*
Chr17:47419917-48261732:+	-11.3	many
Chr6:33053555-33095875:+	-6.6	-
Chr9:19380187-88990615:-	-5.5	many
Chr6:44220937-44221105:-	70.6	*HSP90AB1-as*
Chr14:106209355-106237690:+	30.9	-
Chr17:79477278-79477346:+	22.0	*ACTG1-as*
Chr6:29855961-29910763:-	17.2	-
Chr6:32497901-32525963:-	7.4	*HLA-DRB5*
Chr6:31237776-31320721:-	5.9	-
Chr14:106208067-106235831:-	5.8	-
Chr20:57485406-57485456:+	2.85	*GNAS*
Chr14:67708094-69255647:+	0.7	many

Table 5.5: **A circRNA antisense to *HSP90AB1* is the strongest positively associated circRNA with hypoxia score.** Using a L1/L2 penalised linear regression model for the hypoxia gene expression score, with covariates as the circRNA expression levels, the circRNA associated with the hypoxia score were identified, with coefficients as listed. Negative sign indicates association with lower hypoxia score, and positive sign indicates association with greater hypoxia score. The suffix -as refers to the antisense strand to the circRNA. Per convention, confidence intervals for these coefficients are not presented in this table, as these were fit using penalised linear regression, as described in the Methods.

consistently expressed circRNA correlated to the expression levels of *AGO2* mRNA. This analysis revealed that among all 25 circRNA, chr6:44220937-44221105:- showed the strongest statistically significant positive association with *AGO2* expression, with Spearman's $\rho = 0.23$, $p < 10^{-8}$. The correlation between all circRNA and *AGO2* is shown in Figure 5.8 for reference. This circRNA had not yet been reported in the literature, but in the ensuing sections, further evidence dissecting its statistical associations with the changes expected in the hypoxia response, is presented.

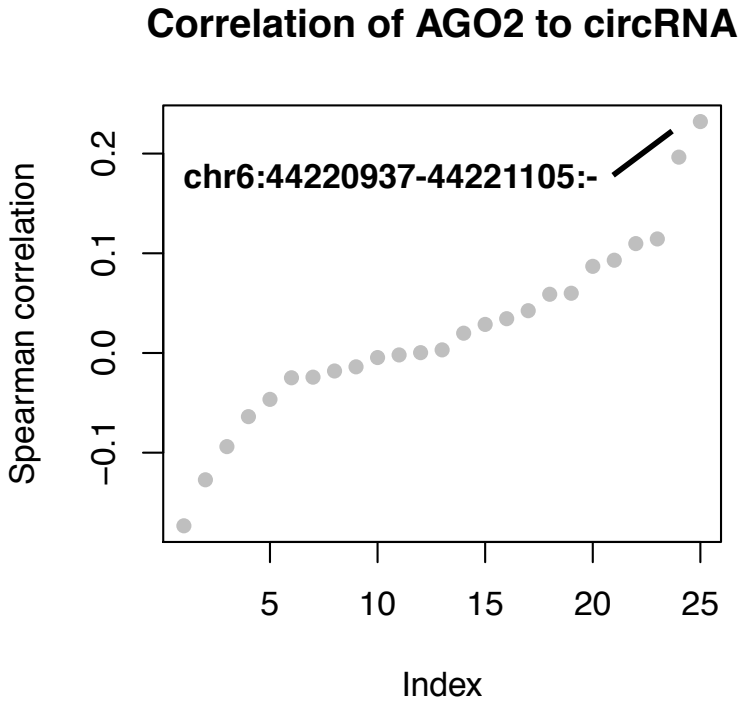

Figure 5.8: **Associations of circRNA expression to *AGO2* expression.** Spearman correlation for the log2 expression of each circRNA to *AGO2* expression is plotted. The circRNA for which the highest correlation was observed was chr6:44220937-44221105:-, antisense to the *HSP90AB1* gene.

5.3.5 Experimental evidence for changing circRNA and miRNA expression patterns in hypoxia

Data produced by the Buffa and Harris labs for the MCF-7 cell line under exposure to hypoxia and normoxia, with subsequent total RNA sequencing, was next considered, and analysed for circRNA expression. Methods describing the experimental and computational design for this dataset are described in the Methods section above.

5.3.6 Composition of circRNA identified by CIRI2.0 and Circ-Seq shows similarity

For the circRNA identified through both computational pipelines, the genomic origins of these species were examined first, and results were compared to the analysis done for the circRNA from the TCGA dataset in Section 5.3.1. The composition was examined by chromosomal location of the circRNA detected, as well as by proportion of circRNA expressed. As depicted in Figure 5.9, the circRNA identified were relatively similar in terms of genomic location, regardless of the pipeline, for polyA minus and ribominus preparations, but were quite different for circRNA in the polyA selected isolates of RNA. Furthermore, as expected, CIRI2 was able to detect a greater quantity of circRNA, to the point where there were clear false positives, such as the detection of putative circRNA on the Y chromosome in the ribominus data and the polyA selected isolates. The comparison of the expression-weighted graphs to non-weighted graphs revealed similarity among the polyA minus and ribominus preparations, which suggested that no chromosome was particularly over-represented in terms of circRNA detection. The analogous finding was not seen in the lowest panel of Figure 5.9, for polyA-selected RNA, potentially due to statistical noise, as fewer circRNA were identified from this isolate.

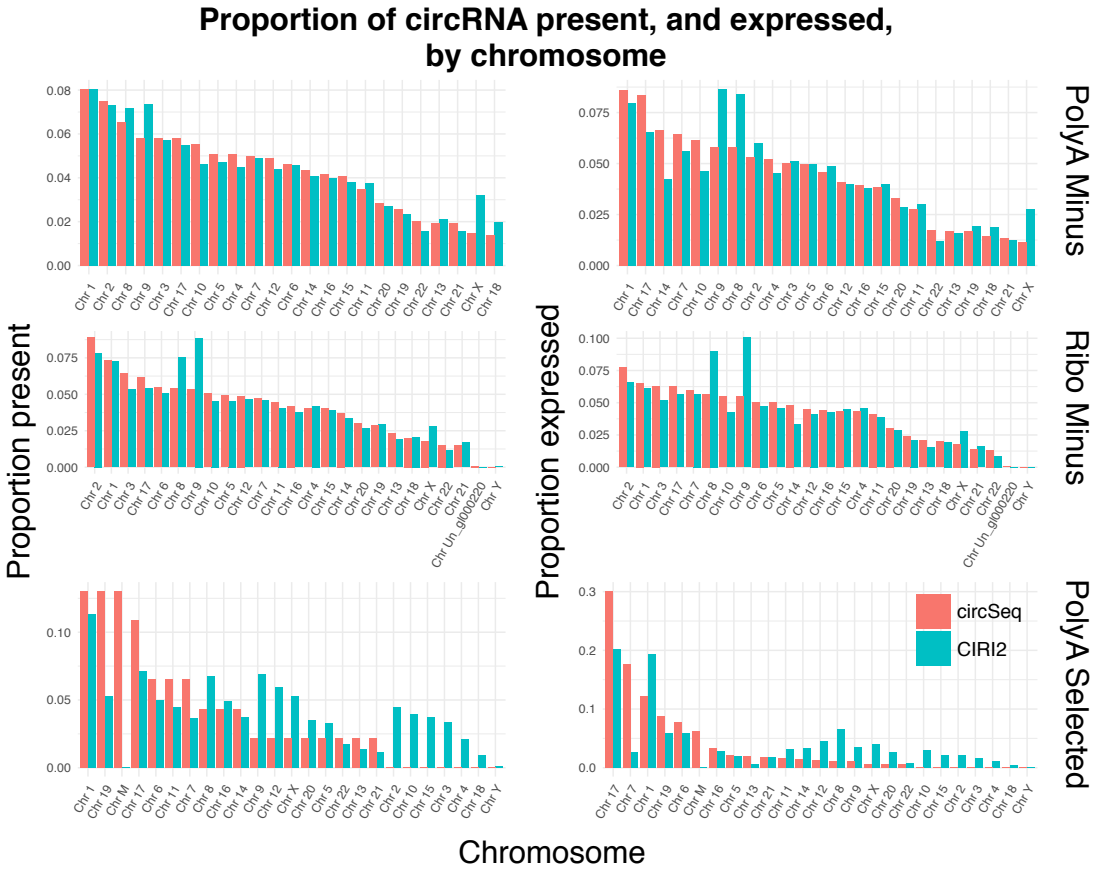

Figure 5.9: **circRNA analysis pipelines showed consistency in circRNA identified, according to genomic location.** Bar graphs depicting the proportion of circRNA present (left) and weighted by expression (right), by chromosome, for the circRNA identified by each circRNA pipeline and from each method of RNA preparation.

Next, the overlap of the circRNA identified from both pipelines was determined, as they relied on different alignment tools, different algorithms, and different filtration settings to identify reads arising from back-spliced junctions, indicative of putative circRNA. Because the tools yielded slightly different annotations that were likely representative of the same circRNA (i.e. with slightly shifted predicted start and end sites), this difference was accounted for by way of a wobble parameter permissive of a slight difference in start and end sites for the identified circRNA species. That is, the proportion of circRNA shared among the samples was determined, where a shared circRNA was defined as one where for both samples the circRNA had start and end sites within some small wobble range. A range of such wobble parameters were tested, from 1-25 base pairs of difference for start and end sites, and the effect of varying this was examined, in terms of its effect on degree of overlap seen between the circRNA identified using both tools. The results obtained are shown in Figure 5.10, and these revealed that among both tools, different circRNA were identified, but there remained a common core subset identified by both tools, and this reproduced the work of Hansen et al., where the reproducibility of the results from different circRNA identification tools was studied [282]. More specifically, CIRI2 and CircSeq identified 825 common circRNA in the polyA-depleted samples, but this represented just 15% of all circRNA species found by CIRI2, and 76% of species found by CircSeq. For ribominus samples, greater concordance in species identified was observed, where 2153 common species were identified, which represented 71% of all circRNA identified using CIRI2, and 55% of those identified by CircSeq. The smallest number of common species was found for the polyA-selected dataset, where just 20 species were found in common, which represented only 2% of all those found by CIRI2, and 43% of those found by CircSeq. This result highlighted the difficulties in reliably detecting circRNA from this polyA-selected experimental data. As an aside, this analysis also showed that CIRI2.0 allowed for the identification of a greater number of circRNA species, highlighting its potential as a more sensitive tool, although validation of these identified circRNA remains to be done to more reliably assess the ground truth of the circRNA present within the cell.

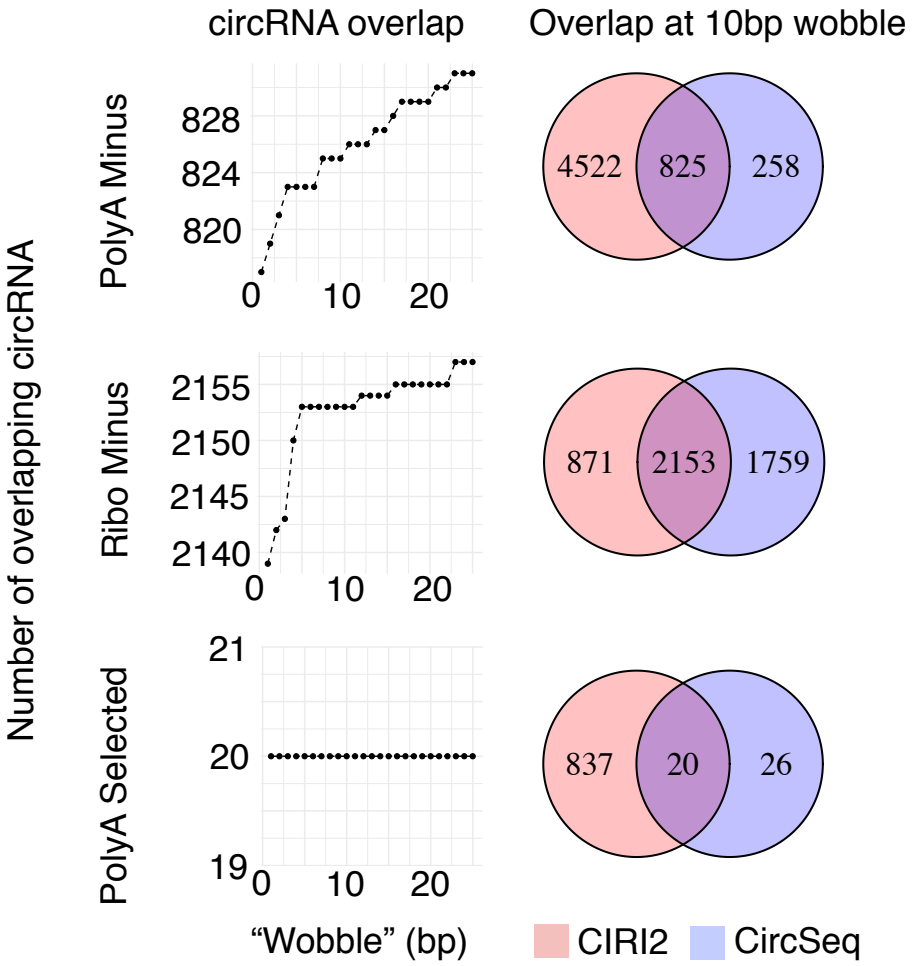

Figure 5.10: **circSeq and CIRI2.0 identify distinct sets of circRNA with strong overlap.** Left: Graphs depicting the number of overlapping circRNA identified by both circSeq and CIRI2.0 pipelines, as a function of wobble parameter, describing the difference allowed between the start sites and end sites of the identified circRNA. Right: Venn diagrams depicting the overlaps between the identified circRNA from the two pipelines, with an allowable wobble of 10bp.

5.3.7 Distinct computational pipelines identify similar circRNA, but those identified from various RNA preparations differ

Having identified the similarities between the results of the circRNA obtained through the various computational pipelines, the differences and similarities between the circRNA identified from the various RNA isolates were examined next. Analysing the sum of the RPM normalised counts across samples (averaged between the two replicates for the polyA-selected data), differences in total circRNA counts were observed, when compared across the different preparatory methods. Specifically, the ribominus preparation isolated the highest total number of circRNA, and the polyA-selected data had the smallest number of circRNA identified, as might be expected. In addition, as shown in Figure 5.11, across all three modes of preparation, and for both computational pipelines, the total count of circRNA was increased in the hypoxic condition, suggesting that these are produced in greater numbers in hypoxia. However, because samples were only in duplicate, statistical testing for the veracity of this difference could not be carried out. Further experimentation is required to truly ascertain these differences.

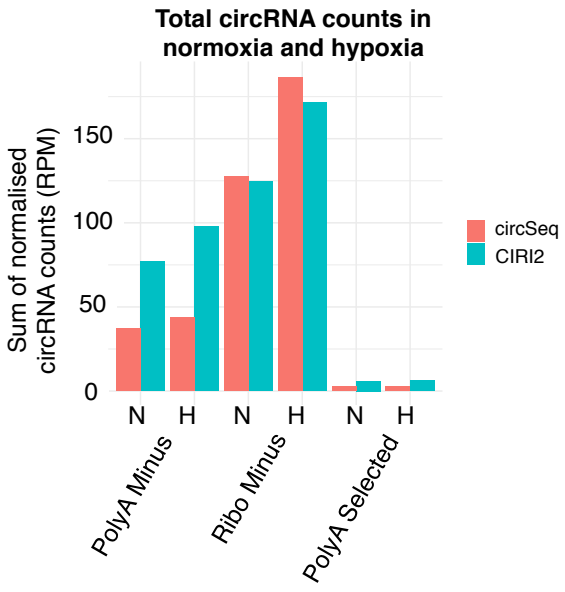

Figure 5.11: **There was an overall increase in circRNA counts in hypoxic samples.** Graph depicting the sum of RPM normalised circRNA counts across each of the isolates of the MCF-7 cell line, showing the differences in counts from each analysis pipeline and under each condition. PolyA selected results are averages of the duplicated conditions. N refers to normoxic conditions, and H refers to hypoxic conditions.

Next, the specific circRNA that changed the most in each sample, either up or down, in normalised count, were identified for the hypoxic condition as compared to the normoxic condition. For each pipeline and isolate, the difference in log2-normalised count was taken as a measure of fold change, and this facilitated the identification of the 10 circRNA in each sample group with the least and greatest fold changes between the hypoxic and normoxic conditions. These results are summarised in Tables 5.6 and 5.7. Different circRNA were identified as having changed to different degrees, depending on the method of isolation used, although between pipelines, results did show low-moderate consistency. The genes overlapping these circRNA molecules were not enriched for any particular pathway, though this analysis with single samples was underpowered to detect these differences, and truly statistically differentially expressed circRNA could not be determined. As an experimental paradigm, this suggests that different methods of isolation yielded different circRNA, and so in order to capture the full response, one may need to use multiple modes of sample preparation.

PolyA Minus, Up

CircSeq		CIRI2.0	
circRNA name	Overlapping genes	circRNA name	Overlapping genes
chr1:117944807-117984947:+	*MAN1A2*	chr2:72945232-72960247:-	*EXOC6B*
chr2:61749745-61761038:-	*XPO1*	chr4:144464662-144465125:+	*SMARCA5*
chr5:95091099-95099324:+	***RHOBTB3***	chr5:95091100-95099324:+	***RHOBTB3***
chr5:167915606-167921655:+	*RARS*	chr6:79752560-79770535:-	*PHIP*
chr7:155465560-155473602:+	***RBM33***	chr6:87925621-87928449:+	*ZNF292*
chr8:124349864-124351686:-	***ATAD2***	chr7:155465561-155473602:+	***RBM33***
chr9:6880011-6893232:+	*KDM4C*	chr8:124349865-124351686:-	***ATAD2***
chr9:86293355-86301070:-	*UBQLN1*	chr9:138773479-138774924:-	*CAMSAP1*
chr12:116668337-116675510:-	*MED13L*	chr14:97299804-97327072:+	*VRK1*
chr14:99924615-99932150:-	*SETD3*	chr20:50342358-50346517:-	*ATP9A*

Ribo-Minus, Up

CircSeq		CIRI2.0	
chr2:9048750-9098771:-	*MBOAT2*	chr4:144464662-144465125:+	*SMARCA5*
chr3:149563797-149639014:+	*RNF13*	chr5:49694941-49707217:-	*EMB*
chr4:37633006-37640126:-	*RELL1*	chr5:179688684-179707608:-	*MAPK9*
chr5:95091099-95099324:+	*RHOBTB3*	chr6:87925621-87928449:+	*ZNF292*
chr6:79752559-79770535:-	*PHIP*	chr7:99621042-99621930:+	***ZKSCAN1***
chr7:99621041-99621930:+	***ZKSCAN1***	chr8:141856359-141900868:-	*PTK2*
chr9:96233422-96261168:+	***FAM120A***	chr9:96233423-96261168:+	***FAM120A***
chr14:99924615-99932150:-	*SETD3*	chr9:138773479-138774924:-	*CAMSAP1*
chr14:102661274-102676199:+	*WDR20*	chr15:80412670-80415142:+	*ZFAND6*
chr20:50273476-50307358:-	*ATP9A*	chr16:4382216-4383520:+	*GLIS2*

polyA selected, Up

CircSeq		CIRI2.0	
circRNA name	Overlapping genes	circRNA name	Overlapping genes
chr1:117944807-117948267:+	***MAN1A2***	chr1:117944808-117948267:+	***MAN1A2***
chr1:117944807-117963271:+	***MAN1A2***	chr1:117944808-117963271:+	***MAN1A2***
chr1:117944807-117984947:+	***MAN1A2***	chr1:117944808-117984947:+	***MAN1A2***
chr6:76388298-76459140:+	***MYO6 SENP6***	chr6:76388299-76459140:+	***MYO6 SENP6***
chr6:151894308-152023140:+	***ESR1 CCDC170***	chr6:151894309-152023140:+	***ESR1 CCDC170***
chr14:96904171-96968937:+	*PAPOLA AK7*	chr6:151894309-151907873:+	***CCDC170***
chr17:58232674-61151375:+	many	chr9:5968019-5988545:-	*KIAA2026*
chr19:34882448-34882947:+	*GPI*	chr11:110124671-110501515:-	*RDX ARHGAP20 FDX1-as*
chr21:36206706-36231875:-	*RUNX1*	chr14:102466326-102500789:+	*DYNC1H1*
chrM:10620-10712:-	-	chr16:25066140-25123317:+	*LCMT1*

Table 5.6: **circRNA identified as increasing in hypoxia vary depending on RNA isolate.** Listing of the top 10 circRNA showing increases in normalised expression in hypoxia vs. normoxia, for each RNA sample and computational pipeline implemented.

PolyA Minus, Up

CircSeq		CIRI2.0	
circRNA name	Overlapping genes	circRNA name	Overlapping genes
chr1:115005725-115007010:-	***TRIM33***	chr1:115005726-115007010:-	***TRIM33***
chr3:150834124-150845771:+	*MED12L*	chr1:155408118-155429689:-	*ASH1L*
chr4:91229394-91234198:+	*CCSER1*	chr6:151932506-151932841:-	*CCDC170-as*
chr6:4891946-4892613:+	*CDYL*	chr7:91924203-91957214:+	*ANKIB1*
chr6:158703294-158735300:+	*TULP4*	chr8:145245687-145255444:+	*HGH1 MROH1*
chr9:88233897-88248289:-	*AGTPBP1*	chr12:46633462-46648719:-	*SLC38A1*
chr9:96233422-96261168:+	*FAM120A*	chr12:97886239-97954825:+	*RMST*
chr10:112723882-112745523:+	*SHOC2*	chr16:80718435-80719026:-	***CDYL2***
chr16:80718434-80719026:-	***CDYL2***	chr20:2944918-2945848:+	*PTPRA*
chr20:46252654-46262380:+	*NCOA3*	chrX:44383248-44386611:-	*FUNDC1*

Ribo-Minus, Up

CircSeq		CIRI2.0	
chr3:63898263-63898901:+	*ATXN7*	chr7:156549087-156619438:-	***LMBR1***
chr5:102432244-102433485:-	*GIN1*	chr8:141615896-141762415:-	*PTK2 AGO2*
chr7:156549086-156619438:-	**LMBR1**	chr9:14639894-14680160:-	*ZDHHC21*
chr8:102570646-102571040:+	*GRHL2*	chr9:33953283-33963789:-	*UBAP2*
chr9:96233422-96238620:+	*FAM120A*	chr9:86294690-86301070:-	*UBQLN1*
chr9:103082546-103111654:-	*TEX10*	chr10:7839010-7844817:+	*ATP5C1*
chr10:128768965-128926028:+	*DOCK1*	chr14:102506573-102507010:+	*DYNC1H1*
chr12:70149163-70150443:+	*RAB3IP*	chr15:25650608-25657118:-	*UBE3A*
chr16:80718434-80719026:-	*CDYL2*	chr17:37864574-37866734:+	***ERBB2***
chr17:37864573-37866734:+	***ERBB2***	chrX:37892787-37893261:+	*SYTL5*

polyA selected, Up

CircSeq		CIRI2.0	
circRNA name	Overlapping genes	circRNA name	Overlapping genes
chr1:117944807-117957453:+	***MAN1A2***	chr1:1586823-1650894:-	*CDK11B/A SLC35E2B MMP23A-as*
chr5:59934576-60053472:-	***DEPDC1B ELOVL7***	chr1:117944808-117957453:+	***MAN1A2***
chr7:26232870-26232993:-	*HNRNPA2B1*	chr1:153895209-154000073:-	many
chr8:99042689-99129618:+	many	chr5:59934577-60053472:-	***DEPDC1B ELOVL7***
chr9:132397698-132397763:-	*ASB6 NTMT1-as*	chr7:158662546-158669382:+	*WDR60*
chr13:34107569-34400023:+	*RFC3 STARD13-as*	chr8:102487410-102505017:+	*GRHL2*
chr16:89973734-89974831:-	*TCF25-as*	chr12:120628101-120785317:-	many
chr17:42275387-42433956:-	**many**	chr14:56078737-56086030:+	*KTN1*
chrM:2823-2882:-	-	chr17:42275388-42433956:-	**many**
chrM:10843-10904:-	-	chr17:57915656-57987972:+	*RPS6KB1 VMP1 MIR21 TUBD1-as*

Table 5.7: **circRNA identified as decreasing in hypoxia vary depending on RNA isolate.** Listing of the top 10 circRNA showing decreases in normalised expression in hypoxia vs. normoxia, for each RNA sample and computational pipeline implemented.

5.3.8 Across samples, there are greater numbers of circRNA identified in the hypoxic condition

The overall expression of all identified circRNA from both pipelines was compared next. As shown in the heatmaps in Figure 5.12, which depict the overall expression of all circRNA identified through both pipelines, across each form of isolation, there were distinct circRNA identified through the different approaches. This again underscored the need for multiple types of isolates in identifying total circRNA content, especially for *de novo* circRNA functional characterisation.

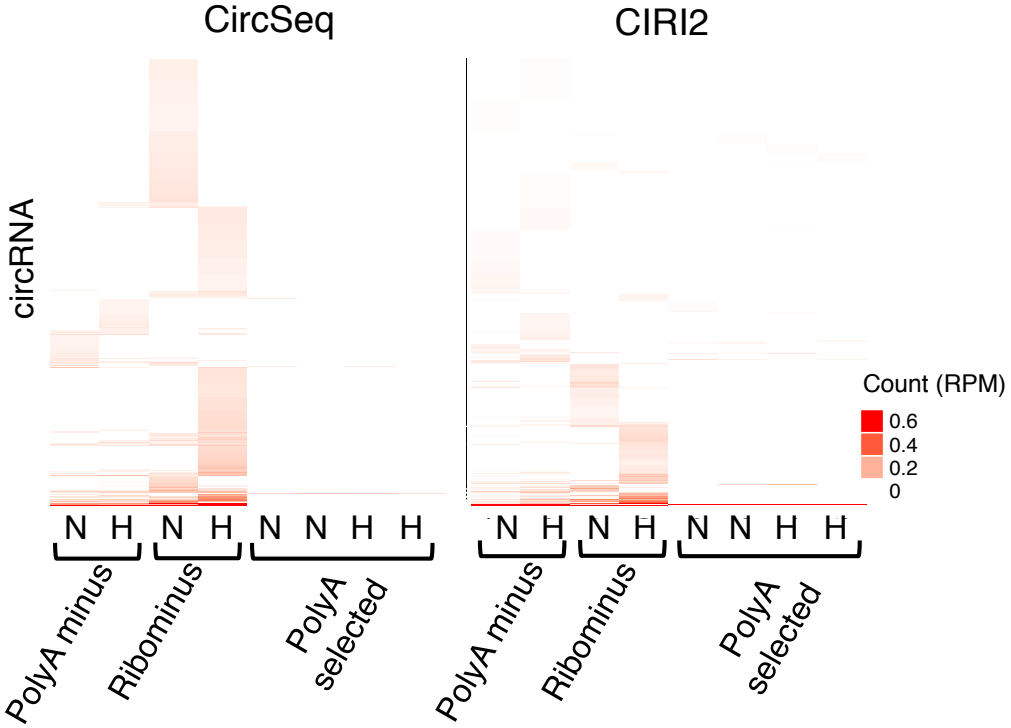

Figure 5.12: **Different circRNA were identified by each informatics pipeline and sample preparation in normoxia and hypoxia.** Heatmaps showing the RPM normalised expression levels for circRNA identified across each sample for the two computational pipelines implemented. N refers to normoxic conditions, and H refers to hypoxic conditions.

5.3.9 A circRNA antisense to *HSP90AB1* is detectable in few samples

Next, in addition to the circRNA which changed most in association with hypoxia, the specific circRNA antisense to *HSP90AB1* was examined in greater detail in the cell line data, as this circRNA was identified as being of potential interest in the analysis of the TCGA dataset, as it showed association with hypoxia and was statistically significantly associated with *AGO2* mRNA expression. Among the circRNA characterised from this cell line, this species was only present in detectable levels in two of the four polyA selected samples (one of each of the sets of two biological replicates), and in no other isolates, and only by the circSeq pipeline. The trend in expression observed was an expression level of 0.05 RPM in the normoxic sample, and an expression level of 0.08 RPM in the hypoxic sample. Given that this increase could only be detected at very low concentrations across just one sample in each condition, statistical significance could not be assessed, and it remains unclear whether this is a true increase in level of this circRNA or whether this is spurious, and any such conclusions would require more thorough and comprehensive experimental validation.

5.4 Discussion

In this chapter, a careful and thorough analysis of a large-scale circRNA dataset was undertaken for matched breast tumour samples, to develop an understanding of the associations between circRNA expression, hypoxia gene signature score, and changes thought to occur in miRNA biogenesis. The common link between these three is the association of hypoxia gene signature score with *AGO2* mRNA expression, and the association of *AGO2* mRNA expression with *HSP90AB1* mRNA expression, raising the hypothesis that it could be stabilised or co-transcribed by an antisense circRNA.

5.4.1 Characterisation of the role of circRNA in hypoxic breast tumour samples

Recently, circularised RNA transcripts have emerged as a further component of the transcriptome, and particular circRNA are thought to play key roles in cancer development and progression. Nair et al., generated a database of circRNA transcripts for the TCGA invasive breast cancer dataset, and showed that these circRNA function as

appropriate predictors to cluster breast cancer samples independently into the well-known molecular subtypes, described by Rouzier et al. in [22]. Using this dataset, the 25 circRNA most commonly expressed circRNA across breast cancer samples were identified, and these were used to identify further statistical relationships using paired miRNA and mRNA expression data from matched tumour samples. In particular, the circRNA most associated with hypoxia gene expression score and prognosis were identified, and in some cases, these showed statistically significant positive correlation with *AGO2* mRNA expression. In doing so, a classifier based on the expression of three circRNA was generated, which was designed to predict survival among breast cancer patients, and it was shown that these circRNA acted independently from tumour subtype, histology, stage, nodal status, and tumour size in a multivariate analysis.

Of the 25 circRNA consistently expressed across tumour samples, the circRNA most strongly associated with hypoxia gene signature score and *AGO2* gene expression was a circRNA antisense to the gene *HSP90AB1*. Further analysis also revealed that this circRNA was co-expressed statistically significantly with the sense gene *HSP90AB1*, and the miRNA hsa-miR-877, which itself has *HSP90AB1* as a predicted target. Furthermore, a study by Rossi highlighted a potential mechanism by which antisense circRNA were able to stabilise their cognate sense transcripts, such as *CDR1as* [283]. The pattern of co-expression of this antisense circRNA led to the hypothesis that this was the case occurring in this scenario, wherein both molecules had statistically significant positive Spearman correlation.

5.4.2 A hypothesised role for antisense *HSP90* circRNA in hypoxia

Based on all of the above results, strong statistical evidence corroborating the association of expression between *HSP90* and *AGO2* has been shown, through analysis of bulk sequencing data from clinical tumour samples. Furthermore, a strong positive association between the *HSP90AB1* mRNA and an antisense circRNA to this gene was shown, and it was also shown that the expression of this antisense circRNA also statistically significantly correlates positively with *AGO2* expression. Furthermore, evidence from the literature for antisense circRNA providing a molecular scaffold, stabilising the corresponding mRNA, was identified, and it was hypothesised that a similar mechanism may explain the interaction of this circRNA and mRNA.

The hypothesis of a functional association between AGO2 and HSP90 in hypoxia is further strengthened by the experimental finding that in hypoxia, AGO2 is hydroxylated on proline residues by the enzyme type I collagen prolyl-4-hydroxylase [284]. Wu et al. showed that this hydroxylation is necessary for the interaction of AGO2 with HSP90; an interaction which then may become more critical in the setting of hypoxia [284].

Thus, it is hypothesised that the hypoxic microenvironment may lead to stabilisation of *HSP90AB1* mRNA through the production of its antisense circRNA, which then enables more AGO2 to localise to stress granules and P-bodies within the cell, to carry out its additional function of miRNA processing, enabling adaptation to a hypoxic, *DICER*-depleted environment. The negative regulation of this feedback loop may then be accomplished by the miRNA miR-877, for which evidence of co-expression with this circRNA has been shown. This miRNA is then predicted to target *HSP90AB1* itself, and may therefore function as a negative feedback regulating the production of P-bodies and enabling miRNA repression in hypoxia.

This network motif hypothesised to occur bears similarity to a validated negative feedback loop experimentally observed involving *DICER1* and the miRNA hsa-let-7, observed under normal circumstances, where it has been shown that global miRNA levels are related to the quantity of expressed hsa-let-7 [285]. Tokumaru et al. showed experimental evidence for this global regulation, and provided direct evidence for hsa-let-7 in reducing *DICER1* levels, suggesting that it may indeed function as an overall negative feedback-based regulator of global miRNA biogenesis [285]. The feedback loop hypothesised in this chapter, summarised in Figure 5.13, may function as an analogue to this in the hypoxic microenvironment, where there may be DICER-independent miRNA biogenesis, though experimental validation would be required for this to be concluded. The data presented supporting this are associative statistical evidence only, and require further study to ensure validity.

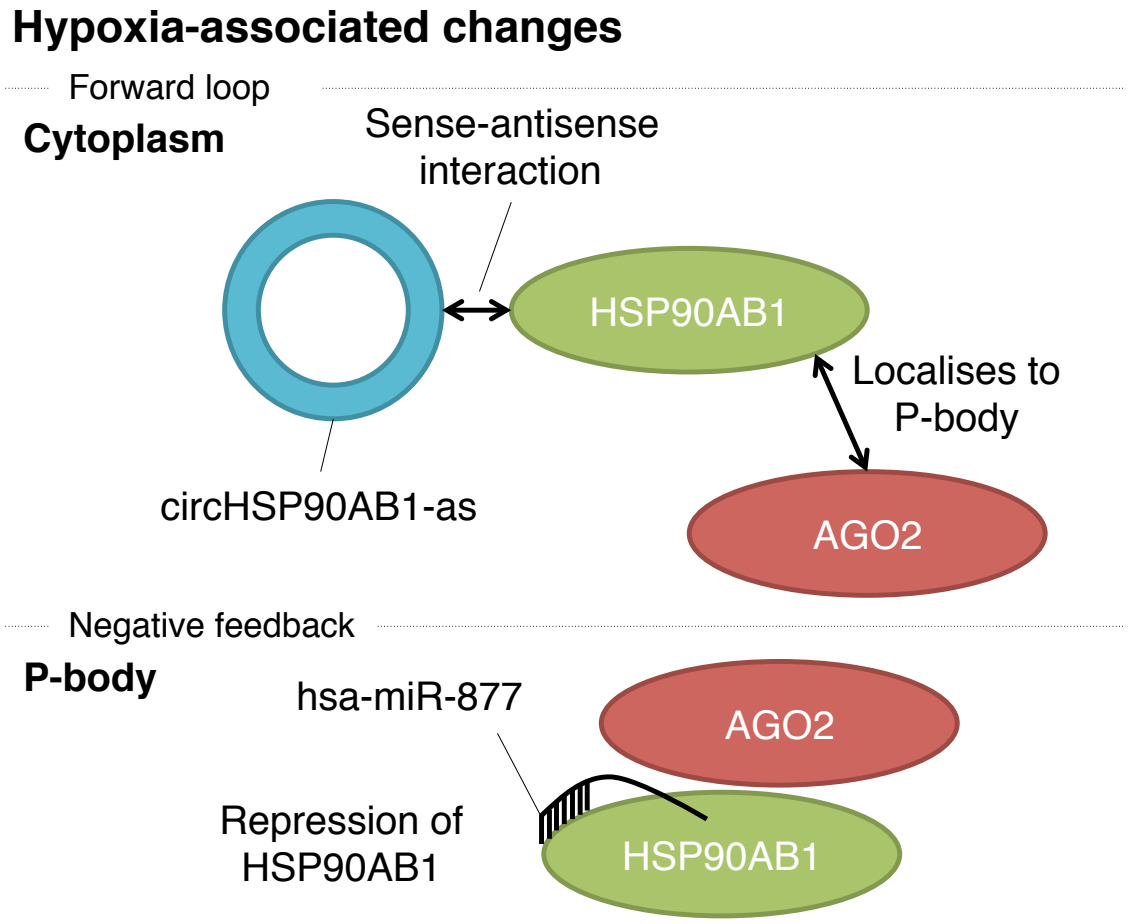

Figure 5.13: **Summary figure describing possible circRNA feedback loop.** Upper panel depicts the circRNA antisense to *HSP90AB1*, correlated with an increase in hypoxia gene expression, potentially stabilising *HSP90AB1*, which may enable its localisation to the P-body for enhanced activity. Lower panel depicts possible negative feedback on the *HSP90AB1* gene, through potential repression by hsa-miR-877, which may act to repress *HSP90AB1* in the P-body.

5.4.3 circRNA characterised from MCF-7 breast cancer cells differ from those found in clinical tumour samples

In this chapter, circRNA expression in a cell line exposed to hypoxia was examined to gain an understanding of how the expression of these species changes in cell lines versus the changes observed in clinical samples. The circRNA up- and down-regulated among MCF-7 cells exposed to 24h of 1% oxygen in a hypoxia chamber were characterised in various modes of sample preparation, wherein samples were prepared with polyadenylated transcripts removed, ribosomal RNA transcripts removed, or polyadenylated transcripts selected, and two distinct circRNA computational pipelines were implemented. These results showed large differences in the quantity and nature of the circRNA identified from each approach, with preparations involving ribosomal RNA depleted showing higher circRNA counts in both normoxia and hypoxia, as well as more unique circRNA identified. Moreover, regardless of the preparation, hypoxic samples showed greater total levels of circRNA, though the study was underpowered to determine whether this difference was statistically significant. This may be indicative of functionality in hypoxia for these RNA molecules, or may suggest differential usage of the splicing machinery in these conditions, as circRNA are thought to compete for splicing machinery in their biogenesis pathway [286], but any firm conclusions regarding this require experimental validation.

5.4.3.1 circRNA identified under different RNA extraction protocols differ greatly

The cell line data analysed in this chapter was derived from a single breast cancer cell line, MCF-7, with RNA prepared in three different ways prior to sequencing, and analysed using two distinct circRNA identification pipelines. Similar to the analysis and results in Hansen et al. in [282], it was shown that the circRNA identified depended largely on the preparation of the RNA sample prior to sequencing but did not appear to depend on the computational pipeline used. Hansen et al. compared the circRNA identified though RNAse-R depleted (ribosomal RNA depleted) and non-depleted samples, and observed that unique circRNA were identified in both cases [282]. With respect to experimental design, the implications of this observation are that it may be necessary to design specific primers to better isolate particular circRNA of interest, or sequence different preparations of RNA from the same sample for a more complete picture of the circRNA present.

5.4.3.2 circRNA identified from cell line experiments differ from clinical specimens

Comparing the results obtained for the analysis of MCF-7 cell line data to those from a re-analysis of the dataset published by Nair et al. for the TCGA breast cancer cohort, revealed important differences. First, the number of circRNA identified differed greatly, and these differences may have been due to the preparation used, as TCGA clinical specimens had RNA isolated using polyA selection, and thus these only contained unmapped junctional reads for the circRNA that were able to survive this stringent selection step. Interestingly, the circRNA antisense to *HSP90AB1*, which showed statistical association with hypoxia gene signature score and association with *AGO2* mRNA expression, was only identified in the polyA-selected isolates for the MCF-7 RNA in normoxia and hypoxia as well, and showed an increase in the hypoxic condition, though statistical significance of this change could not be ascertained due to lack of samples. Notably, this circRNA is not known to polyadenylated, but it may be that there are aspects of the preparation that facilitate its elution in the polyA-selection step; such as a particularly A-rich part of its sequence, however this requires direct experimental evidence to ascertain.

The same degree of increase for this circRNA in hypoxia was not observed in the cell line data as compared to clinical samples, and this difference may have arisen due to a number of factors, related to both the technical aspects of sample preparation, and the experimental conditions used. The circRNA under experimental conditions were isolated after just 24 hours of constant hypoxic conditions, whereas the circRNA in tumour specimens were identified after clinical biopsy, and almost certainly involved a longer and more varying course of hypoxia, potentially encouraging different adaptive mechanisms. Thus, although a statistically significant difference in the levels of the circRNA after 24h of exposure in MCF-7 was not able to be captured, the small initial difference measured suggests that there may be detectable differences within the circRNA transcriptome after hypoxia exposure, and these findings provide further impetus for the generation of experimental evidence.

5.4.4 Experimental investigation of the role of circRNA in hypoxia is underway

In collaboration with Dr. Shaunna Beedie of the Harris lab, based on the hypotheses generated through the work presented in this chapter, validation experiments for the

role of circRNA in hypoxia have been started. Specifically, the aim of the work underway is to characterise the circRNA detected using an array-based capture method, for the HCC-1806 breast cancer cell line under conditions of hypoxia, *AGO2* knockdown, and *DICER1* knockdown. Using this data, validation in a cell line based model for the hypothesised negative feedback loop involving the circRNA antisense to *HSP90AB1* is sought, as well as evidence for its potential role in the hypoxic response, particularly with respect to promoting DICER1-independent AGO2-mediated miRNA biogenesis.

Through this experiment, the goal is to not only validate findings with respect to the circRNA antisense to *HSP90AB1*, but also to contribute to the literature on the reproducibility of circRNA assays through multiple modes of identification, as circRNA identified through both array-based and RNA-sequencing approaches will be compared.

5.5 Summary and Conclusions

In this chapter, the role of circRNA in propagating the hypoxic response was explored, through the identification of a circRNA antisense to *HSP90AB1* and its association with hypoxia gene signature score and *AGO2* mRNA expression. Evidence was shown for how this circRNA may participate as a part of a hypothesised negative feedback loop, mediating the changes in miRNA biogenesis occurring in the hypoxic milieu, for which validation experiments are underway. Experimental data from a cell line characterised under three different RNA preparations was used, and analysed with two different informatics pipelines to examine how various preparations and computational approaches affected the identified circRNA and reproducibility of findings. It was shown that the computational pipelines showed similar results, but the different preparations impacted the species of circRNA identified, providing important information for future experimental approaches to take sample preparation into account. Lastly, the experimental data used from the MCF-7 cell line provided further support to the hypothesised increase in expression for the circRNA antisense to *HSP90AB1*, but due to a lack of samples showing expression of this species, statistical significance of this increase could not be ascertained.

Chapter 6

Endogenous miRNA sponges mediate the generation of oscillatory dynamics for a non-coding RNA network

Abstract

ncRNA have been hypothesised to interact with the transcriptome in a variety of ways within the cell, and recently the theory of ceRNA (competing endogenous RNA) has gained traction in defining key functional roles for various species of ncRNA. This is based on the observation that binding sites for miRNA exist on multiple coding and non-coding transcripts, often in replicate, potentially adding an additional layer of regulation through competitive binding of circulating miRNA, analogous to a miRNA sponge. In addition, recent work has shown how miRNA selectively participate in specific configurations of feedback loops with other species of RNA and RNA binding proteins within the transcriptome, particularly those involving transcription factors, of which many are involved in both carcinogenesis and development. These network structures have been shown to add bistable switch-like behaviour and robustness of dynamics to the intracellular concentration of transcription factors, suggesting mechanisms by which miRNA may act to fine-tune gene expression. The control of time-limited oscillatory behaviour during processes such as generation of neural tissues in the embryo is critically important, and knockdown of ncRNA has been shown to impair these aspects of neural development. Thus, to study this, I consider a model of a non-coding RNA network, with the addition of a miRNA sponge, and show that the action of miRNA sponges on networks of this form acts as a destabilising force. In fact, this adds a further level of control, potentially switching network behaviour from globally stable to oscillatory in the steady state. Analysing a stochastic version of the model, I show that the dynamics can become oscillatory in regions of the parameter space in which steady state behaviour is expected in the deterministic model. These results provide novel hypotheses for the different roles of specific species of miRNA sponges and also a potential experimental paradigm by which to ascribe functional roles to these heretofore uncharacterised RNA.

6.1 Introduction

6.1.1 Network motifs

RNA molecules exist within a tightly controlled, regionalised, admixture within a cell, with each species interacting only with certain others, in broadly activating or inhibitory ways [287]. This activity can be conceptualised as an interaction network, or a directed graph, with each RNA species represented as a node, and edges between interacting species [287]. The pattern of these nodes and the edges between them can be thought of as a graph theoretic, abstract representation of the human transcriptome [287, 288, 289]. It has been shown that the topology, or 'wiring' of this network has a direct link to the time dynamics of the RNA species within a cell that can be observed, and therefore the steady-state behaviour of this RNA species [289, 290, 291].

Recently, this network approach to conceptualising the transcriptome has been extended to include many species of non-coding RNA, such as miRNA, which have inhibitory edges directed towards their predicted mRNA targets [292, 293, 294]. A fundamental question that can then be asked of this network is whether there are sub-graphs of this overall large network that occur more frequently than is expected due to chance alone. Large scale studies, in which this transcriptomic network could be reproduced, have analysed exactly this question, and have discovered over-representation of particular network patterns of interaction, which are termed conserved motifs [295]. Tsang and colleagues identified these network motifs by developing a statistical resampling-based methodology to assess the involvement of a miRNA and its putative targets among a set of feedback and feedforward network motifs, showing enrichment for particular classes of these motifs across the human and mouse transcriptomes [295]. These network motifs comprise a large part of the miRNA - mRNA interactions in the full network, and comprise interactions spanning genes involved across a host of biological processes, including development and neoplastic disease [296]. For instance, the oncogenic miRNA miR-17/92 family and the *MYC* driver gene have been shown to interact with the transcription factor E2F through a conserved network motif, and a similar interaction exists between miR-200 and *ZEB1*, in the epithelial-mesenchymal transition [297, 298, 299]. In fact, even in networks reconstructed from cancer datasets, these sub-networks have been validated for critically involved genes, such as a positive feedback involving p53 and miR-192 [300]. Conserved network motifs, in addition to capturing biological understanding of gene regulation may also lead towards more general strategies for efficient targetting of aberrant pathways driving cancer.

In addition to miRNA, the impact of other species of ncRNA on these networks remains to be determined, as predicted interactions remain elusive for circRNA, lncRNA, and pseudogenes [294, 301, 302, 303]. A common hypothesis for these species of ncRNA is that they may contain elements within their sequence to bind miRNA, so that the miRNA are unable to bind mRNA [304]. This competition for miRNA binding is termed 'sponging', and is thought to be a primary function of certain circRNA, pseudogenes, expressed 3' UTRs, and potentially a function for lncRNA as well [301]. As a part of this transcriptome network, miRNA sponging may be thought of as an inhibitory edge between the sponge RNA molecule and the miRNA [294]. The impacts of these types of edges on miRNA dynamics remain uncharacterised, and in this chapter, I provide the theoretical analysis for an example of such a system.

6.1.2 Modelling network dynamics

The network structure described above gives the rules for the interactions between the species inside the cell. To determine the time-course of such a system, however, it is imperative to consider the dynamics represented within each of the edges between the species. Biological systems have been modelled in a multitude of ways on networks, including Boolean logic and ordinary differential equation (ODE) and partial differential equation (PDE)-based methods; a review of which was written by Hasty et al. in 2001 in [305].

Boolean logic networks encode the state of each network as a Boolean variable, 1 or 0, dependent upon whether a species is active/present, or not [305, 306]. This simplification works very well for gaining an understanding of how network dynamics may look in the steady state, with minimal assumptions [306]. However, owing to its oversimplified nature, reducing time dynamics into Boolean variables loses a great deal of information and nuance for the species, and transient behaviours cannot be described. A particularly salient feature of Boolean network based modelling is the presence of an absorbing state. For a finite number of nodes k, there are 2^k states possible for the entire network. Thus, as the number of time steps approaches infinity, necessarily the states repeat themselves, and this pattern of repeating states in the large time limit is known as the absorbing state. This is a particularly useful feature of Boolean networks to study, because it captures the range of steady states of the network [307]. Inferring the absorbing states of a Boolean network leads to an understanding of which elements of a genetic regulatory network may be co-expressed, or inversely expressed. As such, an analysis of these network states and the involved

nodes may lead to the identification of targets for novel therapeutics leading to synthetic lethality for the dysregulated pathways driving the cancer cell.

Differential equation based methods for studying time-varying behaviour on networks have been widely used throughout biological realms, and have greatly furthered the understanding of how time dynamics create biologically important states [308, 309]. These methods rely on more information than network structure, and therefore intrinsically contain more assumptions than Boolean network models. However, recent work has begun the arduous task of characterising many of these interactions and the functional forms for the dynamics underlying each of the network edges in more explicit terms [310, 311]. Despite the limited experimental evidence, there are general principles that govern the 'rules' followed by network dynamics as modelled through a first-principles ODE approach, summarised as follows. For enzyme kinetics, activation and deactivation follow a generalised functional form based on Hill function kinetics [312]. Secondly, unless otherwise specified, the kinetics of interacting species are assumed to be first-order mass action.

Using these two principles, an approximate ODE description for nearly any biological network can be determined [313]. From this initial description, an analysis of the behaviours and dependence on parameter values can give insight into the possible behaviours of the system. This also can be used to identify biologically relevant scenarios, and has even been used to devise novel hypotheses for biological validation and testing [313]. For instance, Goldbeter predicted circadian rhythm oscillatory behaviour through a theoretical analysis of the circadian clock period protein in Drosophila [314].

A feature of these ODE-based methods that has not been mentioned is the presence of delays in the network dynamics [309]. Delay differential equations are a class of dynamical system which describes the time evolution of a system based on a state at a prior time before the current state, and therefore enables events to happen because of prior interactions, adding 'memory' to the system. These systems are of great importance in the realm of theoretical biology, as they provide a means to describe the potential behaviour for systems where the timescales for various interactions in the network may differ greatly. Previous work from the theoretical biology community has shown that these differences may give rise to unique coupling behaviours between interacting biological elements and stability properties [315, 316], further underscoring the importance of accounting for these delays when present when describing biological systems.

6.1.3 Oscillatory behaviour in biological systems

Oscillatory behaviour is often a critical element of the behaviour of biological systems, across scale, species, stages of development, and in health and disease [317, 318, 319]. In developmental biology, oscillatory behaviour has been characterised as an important feature of vertebral biogenesis, in a process termed somitogenesis [320, 321, 322, 323, 324]. In this stage of development, the cells of the developing organism display synchronised oscillatory behaviour, which results in the development of vertebrae in a coordinated, highly regulated process. For these oscillations to occur, a seminal work in mathematical biology has hypothesised the 'clock and wavefront' model, which predicts the emergence of these oscillations arising from a biochemical network and diffusive effects [325, 326, 327, 328].

Further, in the brain of organisms exhibiting circadian rhythms, patterns of neurotransmitter and neurohormonal release are coupled to oscillatory modes [329]. Oscillations within this system are typically entrained by a system of highly coupled synchronised neuronal populations [329, 330, 331]. In the human brain, this has been localised to the suprachiasmatic nucleus, a neuronal subpopulation located within the hypothalamus. In debilitating states of disease, such as circadian sleep rhythm disorder, these oscillations are thought to be dysregulated, either through exogenous influence, or through disordered coupling between the neurons of the suprachiasmatic nucleus [329, 332]. Mathematical modelling of this subpopulation of coupled oscillators has shown the importance and influence of the coupling strength between these neurons in creating and sustaining these stable oscillations [319, 330, 331].

6.1.3.1 Mechanisms of achieving oscillatory behaviour

As discussed above, the generation of oscillatory behaviour is a critical element of a number of biological systems. Mathematical modelling has shown a number of ways in which these oscillations may be generated (for a review, consult [333]), and in this section, I summarise these methods, with example systems exploiting these properties to generate oscillations in Table 6.1.

The role of coupling between cells is an important one, but here it is posited that in general, coupling is critical for the maintenance and robustness of oscillations generated within biological systems, which is necessary in the face of real-world noise [340, 341, 342]. However, it has been shown that the role of coupling is generally one that is necessary, but not sufficient for the development of oscillations in biological systems [333]. Certainly, without coupling, a single cell oscillator is subject

Characteristic	System examples	References
Negative feedback loops	Circadian rhythms (*PER*, *TIM*, *CLOCK*)	[314]
Time delays	P53/MDM2 signalling, NF-κB signalling	[334, 335, 336]
Sufficient system nonlinearity	Signalling (cAMP, *ERK2*)	[337, 338]
Balancing timescales of opposing reactions	Goodwin oscillator (generalised enzymatic regulation)	[339]

Table 6.1: **Characteristics facilitating oscillatory behaviour in biological systems.** A summary of the four ways proposed by Novák and Tyson in [333] for achieving oscillatory dynamics in gene regulatory networks. Table is adapted from [333]. Note also that many of these examples rely on more than one of these characteristics to achieve oscillatory behaviour.

to a great deal of noise, and the output of the oscillator will be drowned out by the behaviour of other, non-coupled cells in any multicellular organism [340].

6.1.4 Overview of non-coding RNA identified as potential miRNA sponges

The revolution in RNA sequencing technology over the last 5 years has led to the isolation of a diverse population of different RNA species. This diversity is primarily manifested in the discovery of multiple classes of novel non-coding RNAs, many of which have an unknown function. One of the roles of these non-coding RNAs, subject to a great deal of controversy, is miRNA sponging, often termed competitive miRNA binding, wherein a bound miRNA can no longer function to repress its mRNA targets [343]. The controversy over this issue of miRNA sponging stems from largely a physical constraint on the relative concentrations of miRNA and their predicted sponges in the cytoplasmic compartment of the cell [344]. For example, if both relative concentrations are too low, then it is unlikely that the predicted sponges will have any effect, and estimates for the concentrations of these species across different cell types is lacking [344]. Furthermore, while the idea of a miRNA sponge is a compelling one, limited empiric evidence exists for the occurrence of this mode of regulation (repression of the mRNA repressors). The strongest evidence for miRNA sponging comes from the concept of competing mRNA, termed ceRNA, which may competitively inhibit miRNA from binding other targets [345]. For example, the pseudogene *PTENP1*, a pseudogene for *PTEN*, was among the first pseudogenes shown to have a tumour suppressive role, by being targeted by miRNA that would instead target the *PTEN* transcript itself [346]. In addition to this, there is further evidence

for the roles of certain lncRNA, such as that of the H19 lncRNA in sponging the miRNA let-7, which was shown to control the process of differentiation in muscle cells [347].

In addition to ceRNA and lncRNA, other species of non-coding RNA are also thought to function as miRNA sponges. In particularly, as summarised by Thomson and Dinger in their review of the subject, there are 5 predicted species of RNA thought to possibly function as miRNA sponges: ceRNA, lncRNA, 3' UTRs, circRNA, and pseudogenes [301]. While Thomson and Dinger discuss the evidence for and against each of these species, the larger issue of why so many types of miRNA sponges exist within a cell remains unexplored. That is, if sponging is important to biological phenomena, certainly evolutionary selective pressures would have played a role in the optimisation of these sponges. One solution to this apparent paradox is the possibility for different modes of regulation for each of these species of non-coding RNA. For 3' UTRs, circRNA and pseudogenes, there is limited understanding of the regulation and biogenesis of these species, but there is increasing understanding of the regulation of ceRNA (mRNA), and lncRNA. These various species of RNA are likely regulated in very different ways, and therefore may contribute differing dynamics to the networks in which they act.

6.1.5 Research questions

As was have shown in the previous chapters of this thesis, ncRNA such as miRNA and circRNA are likely to interact with coding RNA in networks to produce phenotypic consequences in cancer cells. In this chapter, I seek to extend this understanding to the level of network dynamics, and define and analyse a mathematical model for a common feedback motif. Using this mathematical model, I investigate the conditions under which oscillatory behaviour may arise, and how different non-coding RNA species, inhabiting different regimes of parameter space of the mathematical model, might be utilised by the cell to produce distinct network dynamics.

The remainder of this chapter is structured as follows. In Section 6.2.1 I derive a mathematical model describing system behaviour for a model ncRNA feedback network. Then, in Sections 6.2.2 - 6.2.3, existence and uniqueness of a solution to this dynamical system is proven, steady state behaviour analysis is performed, and conditions for bifurcation are derived. Next, parameter dependence on the steady state is studied, followed by analysis for existence of a bifurcation, and a study of the critical time delays for oscillatory behaviour in Section 6.2.4. Subsequently, in Section 6.2.5, it is shown how time dependence of parameter values may give rise to

dynamically occurring oscillatory behaviour. Lastly, in Section 6.2.6, the stochastic behaviour of the system is studied, revealing the presence of stochastic oscillations.

6.2 Results

6.2.1 Mathematical model definition

For the theoretical study of a ncRNA network with feedback, first consider a feedback motif involving a miRNA and transcription factor that has been shown to occur in the transcriptome more frequently than chance alone would predict. The network structure modelled is analogous to that of the E2F transcription factor and the miR-17/92 oncogenic cluster, well known species in cancer, extended by the addition of a miRNA sponge [297]. That is, the case of a miRNA sponge repressing a miRNA through competitive binding, with the miRNA inhibiting translation of a mRNA for a transcription factor, is considered. The feedback comes in this network as the transcription factor protein functions to increase production of the miRNA which inhibits its own mRNA. Time delays are represented by τ_1 and τ_2 in this system to account for transcription factor mediated activation of transcription, and translation of mRNA into protein, respectively. This feedback motif is depicted graphically in Figure 6.1. The mathematical model is defined as follows, with parameter definitions in Table 6.2. The concentration of sponging RNA over time t is defined as $C(t)$, transcription factor mRNA defined as $F(t)$, transcription factor protein defined as $P(t)$, and miRNA is defined as $M(t)$. Basal rates of production of sponge RNA, miRNA, and transcription factor mRNA are denoted as α_i where $i \in \{C, M, F\}$, respectively and basal rates of degradation of sponge RNA, miRNA, transcription factor mRNA, and transcription factor protein as δ_i with $i \in \{C, M, F, P\}$, respectively. Inhibitory actions between two species i and j are assumed to follow mass-action kinetics (see [348] for a reference), with rate constant denoted k_{ij} for $(i, j) \in \{(C, M), (M, F)\}$ for miRNA sponge repressing miRNA and miRNA repressing transcription factor mRNA, respectively. The rate of production of protein from mRNA for transcription is posited to follow a delayed linear relationship to the amount of mRNA, with an average translation rate of k_P per unit of mRNA. Further, the interaction term between the transcription factor and its back-activation of miRNA production is defined in the following Hill-type function, as done in similar models (e.g. [349]), such that:

$$\alpha_{FM}(P) = \frac{\beta_{FM}}{\left(\frac{\gamma_{FM}}{P}\right)^n + 1}. \tag{6.1}$$

This function describes a sigmoidal relationship defining the increased production of the miRNA species and the concentration of the transcription factor, accounting for both cooperative effects in transcription factor binding to DNA, and saturation behaviour when high concentrations of transcription factor are present.

As a first approximation, because interaction kinetics and production kinetics for the species involved in this system are largely uncharacterised, delayed mass-action kinetics are used for modelling dynamics. Thus, the following equations are obtained, with all derivatives taken with respect to time t, signified by $\dot{C}, \dot{M}, \dot{F}, \dot{P}$ for each of the species as such:

$$\begin{aligned}
\dot{C} &= \alpha_C - \delta_C C - k_{CM} CM \\
\dot{M} &= \alpha_M - \delta_M M - k_{CM} CM - k_{MF} MF + \alpha_{FM}\big(P(t-\tau_1)\big) \\
\dot{F} &= \alpha_F - \delta_F F - k_{MF} MF \\
\dot{P} &= k_P F(t-\tau_2) - \delta_P P.
\end{aligned} \tag{6.2}$$

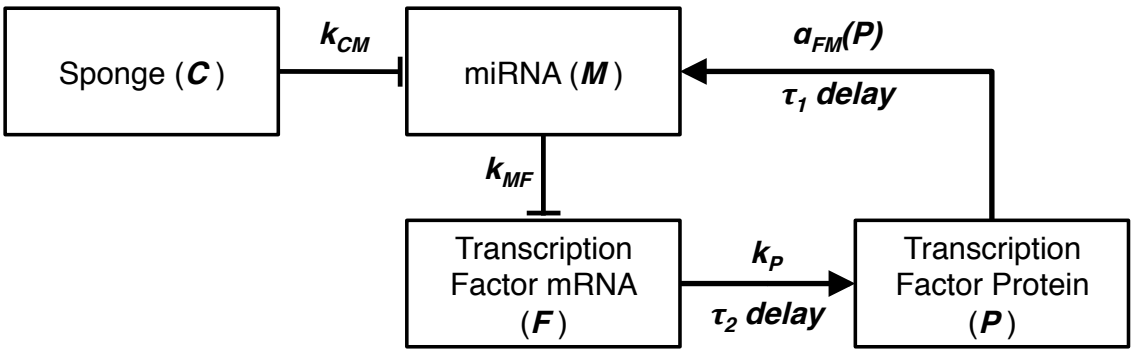

Figure 6.1: **The miRNA sponge network considered.** Directed arrows represent activation-type behaviour, and blunted arrows represent inhibitory behaviour. The system interconnections are overlayed with rate kinetic functions for each of the interactions and time delayed interactions are indicated by τ_1 and τ_2, yielding System 6.2.

6.2.2 Existence and uniqueness of system solution

First, existence and uniqueness of a steady state solution to the ODE system, as defined above in System 6.2 is shown. Any steady state that exists must satisfy the stability condition $\dot{C} = \dot{F} = \dot{M} = \dot{P} = 0$, for some (C, F, M, P). Thus, suppose that there exist a steady state $(C^\star, M^\star, F^\star, P^\star)$ satisfying this. Then it also satisfies the

following:

$$\begin{aligned}
0 &= \alpha_C - \delta_C C - k_{CM} CM, \\
0 &= \alpha_M - \delta_M M - k_{CM} CM - k_{MF} MF + \alpha_{FM}\big(P(t-\tau_1)\big), \\
0 &= \alpha_F - \delta_F F - k_{MF} MF, \text{ and} \\
0 &= k_P F(t-\tau_2) - \delta_P P.
\end{aligned} \tag{6.3}$$

Next, the second expression is rewritten in terms of P only, using the remaining equations to isolate each of other variables in terms of P explicitly. Substituting these expressions into the second expression, a function of P is obtained, defined as $H(P)$ as the following, and at equilibrium, $H(P^\star) = 0$.

$$\begin{aligned}
H(P) = \alpha_M - \alpha_F + \frac{\beta_{FM}\, P^n}{{\gamma_{FM}}^n + P^n} + \frac{\delta_F\, \delta_P\, P}{k_P} - \frac{\delta_M\, k_P\, \left(\alpha_F - \frac{\delta_F\, \delta_P\, P}{k_P}\right)}{\delta_P\, k_{MF}\, P} \\
- \frac{\alpha_C\, k_{CM}\, k_P\, \left(\alpha_F - \frac{\delta_F\, \delta_P\, P}{k_P}\right)}{\delta_P\, k_{MF}\, P\, \left(\delta_C + \frac{k_{CM}\, k_P\, \left(\alpha_F - \frac{\delta_F\, \delta_P\, P}{k_P}\right)}{\delta_P\, k_{MF}\, P}\right)}
\end{aligned} \tag{6.4}$$

Next, it is shown that this steady-state solution $P^\star$ exists and is unique. First, choose strictly positive valued constants C_i such that:

$$H(P) = C_1 + \frac{C_2 P^n}{C_3 + P^n} + C_4 P - \frac{C_5 - C_6 P}{C_7 P} - \frac{C_8 - C_9 P}{C_{10} P + C_{11}}. \tag{6.5}$$

Next, observe:

$$\lim_{P \to 0} H(P) = -\frac{C_5 - C_6 P}{C_7 P} = -\infty, \tag{6.6}$$

since $C_5 > 0$, and

$$\lim_{P \to \infty} H(P) = \lim_{P \to \infty} C_1 + C_2 + C_4 P + \frac{C_6}{C_7} + \frac{C_9}{C_{10}} > 0. \tag{6.7}$$

When differentiating $H(P)$, the first term is a constant, and the following two terms $C_2 P^n/(C_3 + P^n)$ and $C_4 P$ are increasing, and the term $(C_5 - C_6 P)/C_7 P$ is decreasing if and only if (by the quotient rule):

$$-C_6 C_7 P < C_7 (C_5 - C_6 P). \tag{6.8}$$

This implies that

$$C_5 C_7 > 0, \tag{6.9}$$

which is true, since both $C_5, C_7 > 0$ by their definition. Similarly, $\frac{C_8 - C_9 P}{C_{10} P + C_{11}}$ is decreasing if and only if:

$$-C_9(C_{10}P + C_{11}) < (C_8 - C_9 P)C_{10}, \tag{6.10}$$

which is true if and only if

$$-C_9 C_{11} < C_8 C_{10}. \tag{6.11}$$

Then, substituting in the expressions for C_8, C_9, C_{10}, and C_{11}, this inequality is equivalent to:

$$-\alpha_C k_{CM}^2 \delta_f \delta_P k_P \alpha_F < \alpha_C \alpha_F k_{CM} k_P (\delta_P \delta_C k_{MF} - k_{CM} \delta_F \delta_P). \tag{6.12}$$

This simplifies to

$$\delta_C k_{MF} > 0, \tag{6.13}$$

which is true, because both $\delta_C, k_{MF} > 0$, by definition. Thus, $H(P)$ is strictly increasing, and tends to opposing ends of the y axis as $P \to \infty$, and so must have exactly one real root in the interval $(0, \infty)$. That is, there exists a unique steady state solution to the system under consideration. ■

6.2.3 System stability analysis

Now that it has been shown that the system described in System 6.2 has a unique steady state solution at $(C^\star, M^\star, F^\star, P^\star)$, an analysis of stability about the equilibrium solution is performed. To simplify analysis, the system is translated to have equilibrium solution $(0, 0, 0, 0)$, by the following transformation:

$$\bar{C} = C(t - \tau_2) - C^\star, \tag{6.14}$$

$$\bar{M} = M(t - \tau_2) - M^\star, \tag{6.15}$$

$$\bar{F} = F(t - \tau_2) - F^\star, \text{ and} \tag{6.16}$$

$$\bar{P} = P(t) - P^\star. \tag{6.17}$$

Now, the system is re-written using the Taylor series expansion for α_{FM}, about $P^\star$ as

$$\dot{C} = \alpha_C - \delta_C(C + C^\star) - k_{CM}(M + M^\star)(C + C^\star) \tag{6.18}$$

$$\dot{M} = \alpha_M - \delta_M(M + M^\star) - k_{CM}(M + M^\star)(C + C^\star) - k_{MF}(M + M^\star)(F + F^\star) + \sum_{i=0}^{\infty} \frac{\alpha_{FM}^{(i)}(P^\star) P^i (t - \tau)}{i!} \tag{6.19}$$

$$\dot{F} = \alpha_F - \delta_F(F + F^\star) - k_{MF}(M + M^\star)(F + F^\star) \tag{6.20}$$

$$\dot{P} = k_P(F + F^\star) - \delta_P(P + P^\star). \tag{6.21}$$

This simplifies to the following by rearrangement:

$$\dot{C} = -(\delta_C + k_{CM}(M + M^\star))C - k_{CM}MC^\star \tag{6.22}$$

$$\dot{M} = -(\delta_M + k_{CM}(C + C^\star))M - k_{CM}M^\star C - Fk_{MF}(M + M^\star) - k_{MF}MF^\star + \sum_{i=1}^{\infty} \frac{\alpha_{FM}^{(i)}(P^\star)P^i(t-\tau)}{i!} \tag{6.23}$$

$$\dot{F} = -(\delta_F + k_{MF}(M + M^\star))F - k_{MF}MF^\star \tag{6.24}$$

$$\dot{P} = k_P F - \delta_P P \tag{6.25}$$

Linearising the system around the equilibrium point $(0, 0, 0, 0)$ and evaluating the Jacobian, J, yields:

$$J(\lambda) = \begin{pmatrix} -(\delta_C + k_{CM}M^\star) & -k_{CM}C^\star & 0 & 0 \\ -k_{CM}M^\star & -(\delta_M + k_{CM}C^\star + k_{MF}F^\star) & -k_{MF}M^\star & e^{-\lambda\tau}\alpha'_{FM}(P^\star) \\ 0 & -k_{MF}F^\star & -(\delta_F + k_{MF}M^\star) & 0 \\ 0 & 0 & k_P & -\delta_P \end{pmatrix} \tag{6.26}$$

This system is asymptotically stable if and only if all eigenvalues λ have negative real part. Thus, computing $|\lambda I - J(\lambda)|$, the following characteristic polynomial is obtained:

$$\lambda^4 + A_1\lambda^3 + A_2\lambda^2 + A_3\lambda + A_4 + B_1\lambda e^{-\lambda\tau} + B_2 e^{-\lambda\tau} = 0, \tag{6.27}$$

such that:

$$A_1 = \delta_C + \delta_M + \delta_F + \delta_P + k_{CM}(C^\star + M^\star) + k_{MF}(M^\star + F^\star) \tag{6.28}$$

$$\begin{aligned} A_2 =& (\delta_C + k_{CM}M^\star)(\delta_F + k_{MF}M^\star) + (\delta_C + k_{CM}M^\star)(\delta_M + C^\star k_{CM} + F^\star k_{MF}) \\ &+ (\delta_F + k_{MF}M^\star)(\delta_M + C^\star k_{CM} + F^\star k_{MF}) + \delta_P(\delta_C + k_{CM}M^\star) \\ &+ \delta_P(\delta_F + k_{MF}M\star) + \delta_P(\delta_M + C^\star k_{CM} + F^\star k_{MF}) - C^\star k_{CM}^2 M^\star \\ &- F^\star k_{MF}^2 M^\star \end{aligned} \tag{6.29}$$

$$\begin{aligned}A_3 =&(\delta_C + k_{CM}M^\star)(\delta_F + k_{MF}M^\star)(\delta_M + C^\star k_{CM} + F^\star k_{MF})\\ &+ \delta_p(\delta_C + k_{CM}M^\star)(\delta_F + k_{MF}M^\star)\\ &+ \delta_p(\delta_C + k_{CM}M^\star)(\delta_M + C^\star k_{CM} + F^\star k_{MF})\\ &+ \delta_p(\delta_F + k_{MF}M^\star)(\delta_M + C^\star k_{CM} + F^\star k_{MF})\\ &- C^\star\delta_p k_{CM}^2 M^\star - \delta_p F^\star k_{MF}^2 M^\star - C^\star k_{CM}^2 M^\star(\delta_F + k_{MF}M^\star)\\ &- F^\star k_{MF}^2 M^\star(\delta_C + k_{CM}M^\star)\end{aligned} \tag{6.30}$$

$$\begin{aligned}A_4 =&\delta_P(\delta_C + k_{CM}M^\star)(\delta_F + k_{MF}M^\star)(\delta_M + C^\star k_{CM} + F^\star k_{MF})\\ &- C^\star\delta_P k_{CM}^2 M^\star(\delta_F + k_{MF}M^\star) - \delta_P F^\star k_{MF}^2 M^\star(\delta_C + k_{CM}M^\star)\end{aligned} \tag{6.31}$$

$$B_1 = \alpha'_{FM}(P^\star)k_{MF}F^\star k_p \tag{6.32}$$

$$B_2 = \alpha'_{FM}(P^\star)k_{MF}F^\star k_p(\delta_C + k_{CM}M^\star) \tag{6.33}$$

Now, if $\tau = 0$, the characteristic polynomial becomes

$$\lambda^4 + A_1\lambda^3 + A_2\lambda^2 + (A_3 + B_1)\lambda + (A_4 + B_2) = 0 \tag{6.34}$$

In this case, deriving an analytic expression for the roots quickly becomes intractable, but by the Routh-Hurwitz criterion (introduced in [350]), it can be quickly numerically determined whether all roots have negative real part, and if so, the system is globally asymptotically stable.

Bifurcation criteria Next, the case when $\tau \neq 0$ is considered, particularly when the roots of the characteristic polynomial have a non-zero complex part, giving rise to oscillatory behaviour. As τ varies, the roots of the characteristic polynomial may change sign of the real part, and so for a bifurcation to occur, there must exist a root of the characteristic polynomial of the form $\lambda = i\omega$ with $\omega \in \mathbb{R}$ for some value of τ, and then:

$$\omega^4 - iA_1\omega^3 - A_2\omega^2 + iA_3\omega + A_4 + B_1\omega(i\cos(\omega\tau) + \sin(\omega\tau)) + B_2(\cos(\omega\tau) - i\sin(\omega\tau)) = 0. \tag{6.35}$$

Separating out real and imaginary parts,

$$\omega^4 - A_2\omega^2 + A_4 + B_1\omega\sin(\omega\tau) + B_2\cos(\omega\tau) = 0 \tag{6.36}$$

$$-A_1\omega^3 + \omega A_3 + B_1\omega\cos(\omega\tau) - B_2\sin(\omega\tau) = 0 \tag{6.37}$$

Which leads to the equivalent system of equations

$$B_1\omega(\omega^4 - A_2\omega^2 + A_4) + B_2(A_1\omega^3 - A_3\omega) = -\sin(\omega\tau)((B_1\omega)^2 + B_2^2) \tag{6.38}$$

$$B_2(\omega^4 - A_2\omega^2 + A_4) - B_1\omega(A_1\omega^3 - A_3\omega) = -\cos(\omega\tau)((B_1\omega)^2 + B_2^2). \tag{6.39}$$

Squaring both sides and taking the sum yields

$$((B_1\omega)^2 + (B_2)^2)((\omega^4 - A_2\omega^2 + A_4)^2 + (A_1\omega^3 - A_3\omega)^2 - (B_1\omega)^2 - (B_2)^2) = 0. \tag{6.40}$$

This (assuming B_1 and B_2 are non-zero) implies

$$(\omega^4 - A_2\omega^2 + A_4)^2 + (A_1\omega^3 - A_3\omega)^2 - (B_1\omega)^2 - (B_2)^2 = 0, \tag{6.41}$$

which reduces to:

$$\omega^8 + (A_1^2 - 2A_2)\omega^6 + (A_2^2 - 2A_1A_3 + 2A_4)\omega^4 + (A_3^2 - 2A_2A_4 - B_1^2)\omega^2 + (A_4^2 - B_2^2) = 0. \tag{6.42}$$

Letting $z = \omega^2$, $f(z)$ is defined such that:

$$f(z) = z^4 + (A_1^2 - 2A_2)z^3 + (A_2^2 - 2A_1A_3 + 2A_4)z^2 + (A_3^2 - 2A_2A_4 - B_1^2)z + (A_4^2 - B_2^2). \tag{6.43}$$

Now, suppose that there are positive roots $z_1, \ldots, z_k$ for $f(z)$ as defined above, and then take, for each root z_k, $\omega_k = \sqrt{z_k}$. Then the solutions in τ to Equation 6.38 may be defined as τ_k^j such that:

$$\tau_k^j = \frac{1}{\omega_k}\left\{\arcsin\left(-\frac{B_1\omega_k(\omega_k^4 - A_2\omega_k^2 + A_4) + B_2(A_1\omega_k^3 - A_3\omega_k)}{(B_1\omega_k)^2 + B_2^2}\right) + 2\pi j\right\}. \tag{6.44}$$

Define now τ_0 such that:

$$\tau_0 = \tau_{k_0}^{j_0} = \min_{k,j}\tau_k^j. \tag{6.45}$$

Then by the Hopf bifurcation theorem, if such a τ_0 exists, a Hopf bifurcation exists for the values of the total delay $\tau_1 + \tau_2 > \tau_0$.

As a numerical example, consider the system for the following parameter values, chosen because they fall within a realistic range for known range parameters for

mammalian cells as used in similar models (e.g. [335, 351]). The parameter values considered are: $\alpha_C = 1 \text{ mol} \cdot \text{min}^{-1}$, $\delta_C = 0.01 \text{ min}^{-1}$, $\alpha_F = 1 \text{ mol} \cdot \text{min}^{-1}$, $\delta_F = 0.1 \text{ min}^{-1}$, $\alpha_M = 1 \text{ mol} \cdot \text{min}^{-1}$, $\delta_M = 1 \text{ min}^{-1}$, $k_P = 10 \text{ mol} \cdot \text{min}^{-1}$, $\delta_P = 0.1 \text{ min}^{-1}$, $k_{CM} = 10 \text{ min}^{-1} \cdot \text{ mol}^{-1}$, $k_{MF} = 0.1 \text{ min}^{-1} \cdot \text{ mol}^{-1}$, $\beta_{FM} = 200 \text{ mol} \cdot \text{min}^{-1}$, $\gamma_{FM} = 100$ mol, and $n = 8$, with both cases of $\tau_1 = \tau_2 = 0.5$ min and $\tau_1 = \tau_2 = 0.8$ min, depicted in Figure 6.2A and B, respectively. These parameter values give a critical time τ_0 of 1.43 for which if $\tau_1 + \tau_2 > \tau_0$, there is an oscillatory solution, and below which there is a steady state solution, as shown in Figure 6.2A and B.

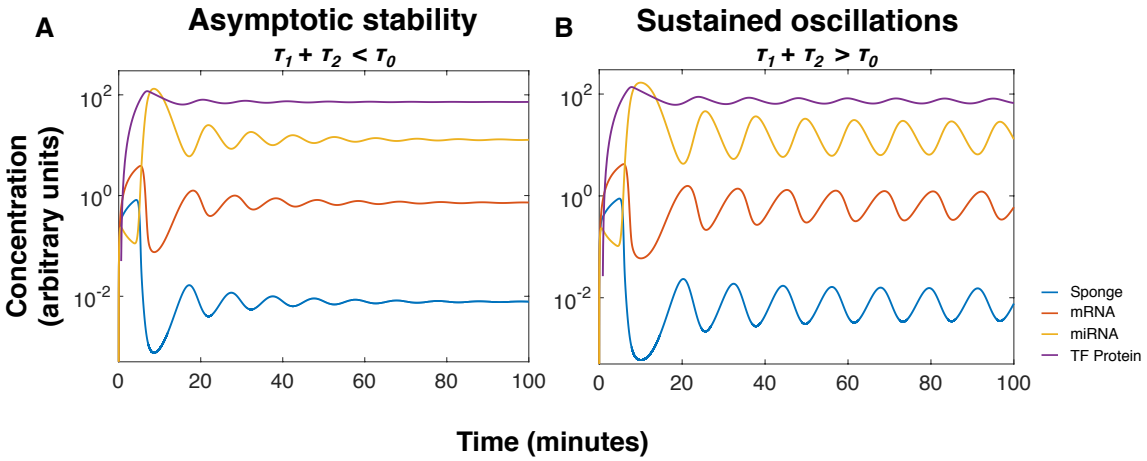

Figure 6.2: **Increasing system delay past critical threshold induces steady oscillatory behaviour, traversing a Hopf bifurcation.** Plots depict that for the same parameter values, the effect of having $\tau_1 + \tau_2$ below (A) and above (B) the critical time threshold τ_0 as derived above, based on the Hopf bifurcation theorem. Parameter values used for this simulation are: $\alpha_C = 1 \text{ mol} \cdot \text{min}^{-1}$, $\delta_C = 0.01 \text{ min}^{-1}$, $\alpha_F = 1 \text{ mol} \cdot \text{min}^{-1}$, $\delta_F = 0.1 \text{ min}^{-1}$, $\alpha_M = 1 \text{ mol} \cdot \text{min}^{-1}$, $\delta_M = 1 \text{ min}^{-1}$, $k_P = 10 \text{ mol} \cdot \text{min}^{-1}$, $\delta_P = 0.1 \text{ min}^{-1}$, $k_{CM} = 10 \text{ min}^{-1} \cdot \text{ mol}^{-1}$, $k_{MF} = 0.1 \text{ min}^{-1} \cdot \text{ mol}^{-1}$, $\beta_{FM} = 200 \text{ mol} \cdot \text{min}^{-1}$, $\gamma_{FM} = 100$ mol, and $n = 8$, with τ_1 and τ_2 indicated as above.

6.2.4 Parameter sensitivity analysis

Because many of the parameters used within the definition of this system remain uncharacterised by experimental evidence, to understand their influence on the long term behaviour of the system, a sensitivity analysis is performed. In order to encompass a wide range of possible parameter values, ranges as outlined in Table 6.2 were considered. The values for these range parameters were specified from global studies of the mRNA and protein components of the mammalian transcriptome, and are meant to capture much of the variation seen in these large-scale studies [351].

Parameter	Description	Values
α_C	Basal production rate of miRNA sponge	{0.1, 1, 5, 10 mol · min^{-1}}
δ_C	Degradation rate of miRNA sponge	{0.0001, 0.001, 0.01, 0.1, 1 min^{-1}}
α_F	Basal production rate of mRNA	{0.1, 1, 5, 10 mol · min^{-1}}
δ_F	Degradation rate of mRNA transcript	{0.0001, 0.001, 0.01, 0.1, 1 min^{-1}}
α_M	Basal production rate of miRNA	{0.1, 1, 5, 10 mol · min^{-1}}
δ_M	Degradation rate of miRNA	{0.0001, 0.001, 0.01, 0.1, 1 min^{-1}}
k_P	Basal production rate of transcription factor protein from mRNA	{0.1, 1, 5, 10 mol · min^{-1}}
δ_P	Degradation rate of transcription factor protein	{0.0001, 0.001, 0.01, 0.1, 1 min^{-1}}
k_{CM}	Rate of binding of miRNA sponge to miRNA	{0.01, 0.1, 1 10, 100 min^{-1} · mol^{-1}}
k_{MF}	Rate of binding of miRNA to mRNA	{0.1, 1, 10, 100, 200 min^{-1} · mol^{-1}}
β_{FM}	Maximum activation of miRNA production by transcription factor	{ 0.1, 1, 10, 100, 200 mol · min^{-1}}
γ_{FM}	Transcription factor concentration for half-maximal activation of miRNA production	{1, 10, 100, 200 mol}
n	Hill coefficient for activation of miRNA production by transcription factor	{2, 4, 8}
τ_1	Delay in activation of transcription of miRNA by activating transcription factor	{0, 0.01, 0.1, 1, 10, 100 min}
τ_2	Delay in translation of transcription factor from mRNA transcript	{0, 0.01, 0.1, 1, 10, 100 min}

Table 6.2: **Description of parameters and associated values as considered for sensitivity analysis.**

Given the large size and high dimensionality of this parameter space, Latin hypercube sampling (LHS) was employed to fairly sample the complete parameter space [352]. Using LHS, the space was sampled, generating 10^5 evenly distributed parameter values encompassing the range of the parameter space, and the system solution was computed for these 10^5 parameter sets. For each of the sampled parameters, the steady state value was determined, as well as whether a bifurcation could exist by the Hopf bifurcation criteria, and if it did exist, the critical time. In this way, the dependence of the steady state on the various parameters, as well as both the existence and critical time τ_0 of the Hopf bifurcation, was determined. To determine the sensitivities on steady state values whilst accounting for the sensitivity to the other parameters, the partial rank correlation coefficient for every parameter on each of the steady state values was considered, relative to all other parameters. In effect, this provided a 'corrected' estimate of the independent influence of each parameter on steady state concentration for the system under consideration. Results of this are depicted for each of the parameters on each of the four steady state values in Figure 6.3.

As expected, rates of production are positively associated with the steady state of each species, and rates of degradation are negatively associated with the steady state of each species. Further, these results show a strong negative dependence for the steady state value of the miRNA sponge on the the kinetic parameter k_{CM}.

Additionally, the parameters controlling the kinetic rate of the feedback activation of miRNA production, β_{FM} and γ_{FM} affect the miRNA and protein levels in predictable ways, with β_{FM} showing a negative effect on miRNA sponge levels, as more miRNA would be produced, reducing the levels of sponges by binding with them. Lastly, there is minimal dependence identified for the steepness of the Hill function controlling the activation kinetics on any of the steady state values, and as expected from the mathematical derivation for the steady state values, τ_1 and τ_2 show no dependence on any of the steady state values.

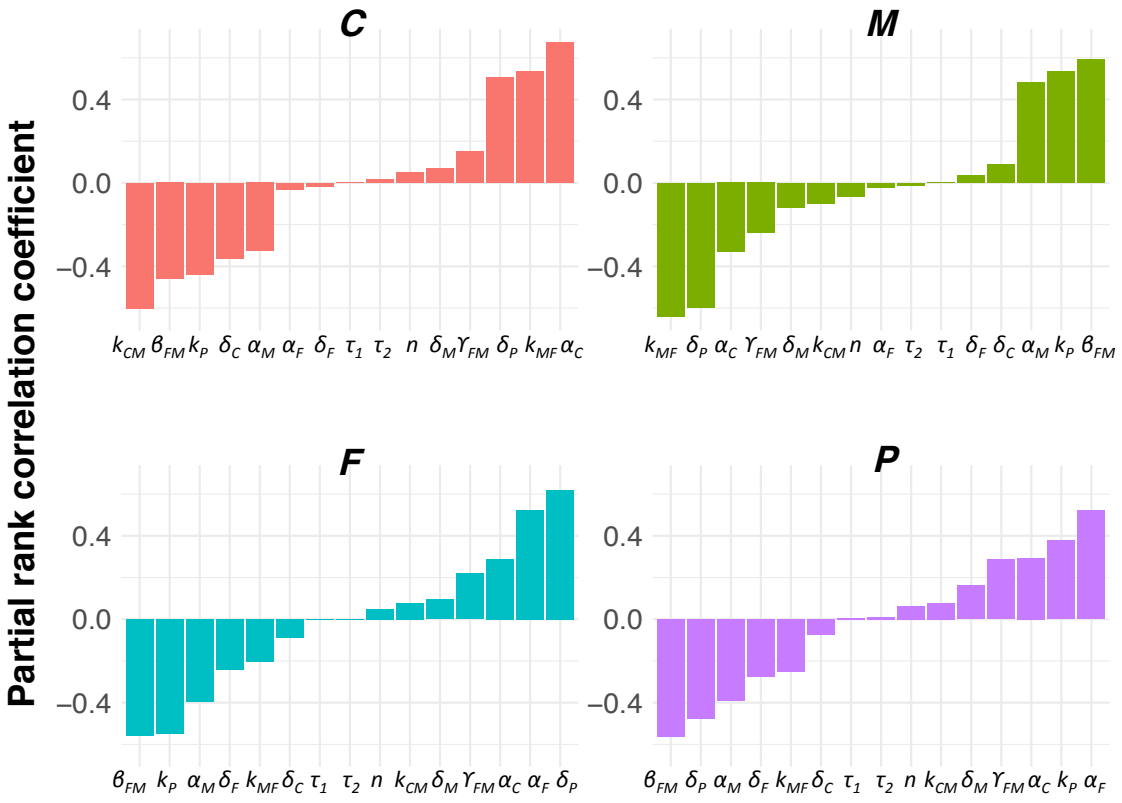

Figure 6.3: **Partial rank correlation coefficients for each parameter value correlated with steady state values for modelled species**. Correlations are taken partial to all other parameter values. Parameter values were sampled using Latin hypercube sampling with 10^5 points from the parameter space, and for each of these parameter combinations, steady state values were computed, from which the partial rank correlation coefficients can be presented. C refers to the steady state concentration of the miRNA sponge, M to the miRNA, F to the transcription factor mRNA, and P to the transcription factor protein.

6.2.4.1 Drivers of overall system behaviour

Analysing the steady state values in combination with each other, by considering the four species as individual components within the four dimensional state space, the sensitivity of each parameter in a global sense was determined. That is, the Euclidean distance from the origin to the steady state through the state space was determined, and it was asked how each of the individual parameters perturbed the distance from this point to the origin, as a measure of overall system behaviour. The results, summarised in Figure 6.4, show that the strongest determinant in a positive sense with the global system behaviour is the degradation rate of protein, and next is the production rate of miRNA sponge. As the protein degrades faster, there is less of a feedback effect producing miRNA, and because there are fewer miRNA, the system is able to achieve greater molecule counts. Likewise, the production of the miRNA sponge also leads to fewer circulating miRNA, which enables the system to also achieve higher concentrations of protein. Further substantiating the importance of miRNA to global system dynamics is the strong negative association between δ_M, the miRNA degradation rate, and the overall state's distance from the origin. As more miRNA are degraded, the system is able to achieve higher overall concentrations of all molecules. Thus, the greatest determinant of overall molecular counts within this system is, in fact, the miRNA-related parameters, suggesting that miRNA dysregulation would act to perturb this system

6.2.4.2 Bifurcation existence

To determine the effect of parameter values on the existence of a bifurcation, the regions within the parameter space where a bifurcation may exist were sought, through LHS of the parameter space. Once the regions were identified, to determine the locations of these regions, the effect of each of the variables on the presence or absence of a bifurcation was determined using a logistic regression model with response variable defined by the presence (1) or absence (2) of a bifurcation. The predictors used in the model were the 15 parameters of the model system, as listed in Table 6.2. To ensure comparability, because only evidence for dependence whilst correcting for all other parameter values was sought, and not linear relationships with parameter value (as such a dependence is certainly nonlinear), the parameters were rank-normalised to values between 0 (least) and 1 (greatest).

Fitting the model using logistic regression, coefficients of the model were obtained, and their associated 95% confidence intervals. While these relationships are not

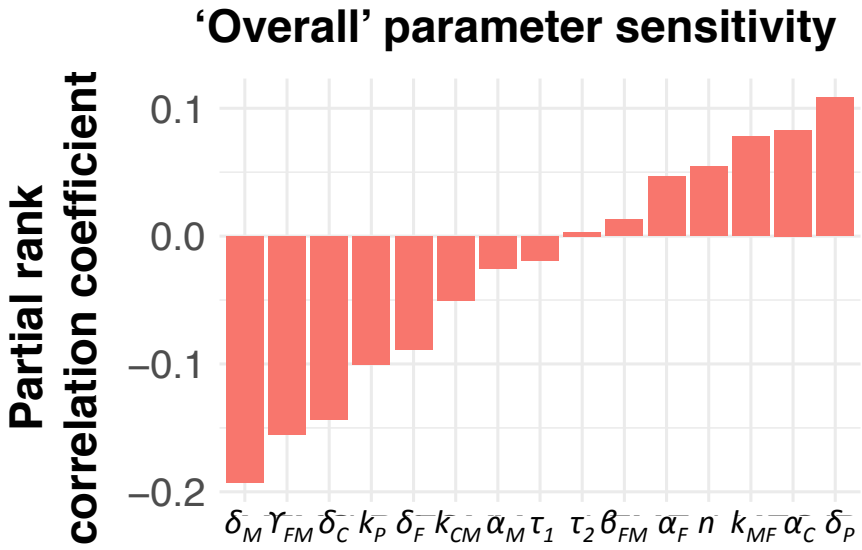

Figure 6.4: **Partial rank correlation coefficients for each parameter value correlated with Euclidean norm of system steady state values.** Correlations are taken partial to all other parameter values. Parameter values were sampled using Latin hypercube sampling with 10^5 points from the parameter space, and for each of these parameter combinations, steady state values were computed, from which the partial rank correlation coefficients can be presented.

expected to be expressly linear (as they are derived based upon a highly nonlinear problem, as defined by the bifurcation criteria), understanding a linearised version of this relationship can provide insight into how the parameter values affect system behaviour. Summarized in Figure 6.5, the model coefficients and associated 95% confidence intervals are depicted on a waterfall plot, showing the dependence of each of the parameter values on the response variable of bifurcation existence. This analysis reveals a strong dependence on the presence of a bifurcation on the values of β_{FM} and γ_{FM}, wherein for lower values (i.e. slower kinetics of miRNA activation), a bifurcation is more likely to occur. Further, there was little dependence of miRNA sponge parameters (α_C, δ_C, and k_{CM}) on the existence of a bifurcation; rather that the miRNA-mRNA feedback loop is the primary determinant of the existence of a bifurcation. Lastly, as expected, the values of τ_1 and τ_2 have little dependence on the existence of a bifurcation, as shown through the derivation in Section 6.2.3.

6.2.4.3 Effect of parameter values on critical time

Next, the biological relevance of each parameter on traversing a bifurcation point of the dynamical system was determined, for a set of parameters where a bifurcation was

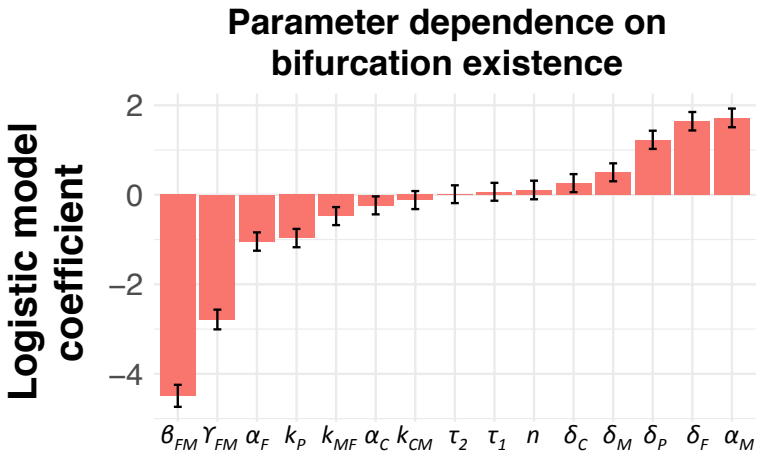

Figure 6.5: **Logistic regression model coefficients for model parameters.** These are depicted with the associated 95% confidence interval for predicted coefficient. Model was trained as a classifier for whether regression would occur or not on 10^5 parameter value combinations sampled from the space of possible values by Latin hypercube sampling. A positive coefficient indicates more likely to associate with existence of a bifurcation, and a negative coefficient indicates more likely to associate with global asymptotic stability.

known to exist. Thus, in addition to examining the dependence of the sampled parameter values on the existence of a bifurcation, for the cases where a bifurcation did exist, the dependence of the critical time on parameter values was determined. First, using the sampled parameter values, the cases in which a bifurcation existed were identified. Then, the partial rank correlation coefficient of each parameter as it correlated with the critical time for the bifurcation was determined, and this uncovered the linearised relationship between critical time and parameter value.

This analysis, summarised in Figure 6.6, shows that in the cases where there is a bifurcation, the largest determinant of the critical time is the degradation rate of the protein, with a higher degradation rate associated with a reduced critical time. This points to yet another manner in which the cell may regulate whether the system oscillates or not, as the regulation of proteases will likely influence this rate. Further, the effect of miRNA sponge parameters is small, yet non-negligible. As might be expected, miRNA sponges have a destabilising impact on the system, with k_{CM}, the kinetic binding coefficient for such sponges, and α_C, the production rate of the sponge species, both associated with reducing the critical time necessary to achieve bifurcation. Additionally, these results reinforce the role of miRNA as a stabilising presence within the system, with the miRNA production and degradation rates both positively associated with the critical bifurcation time.

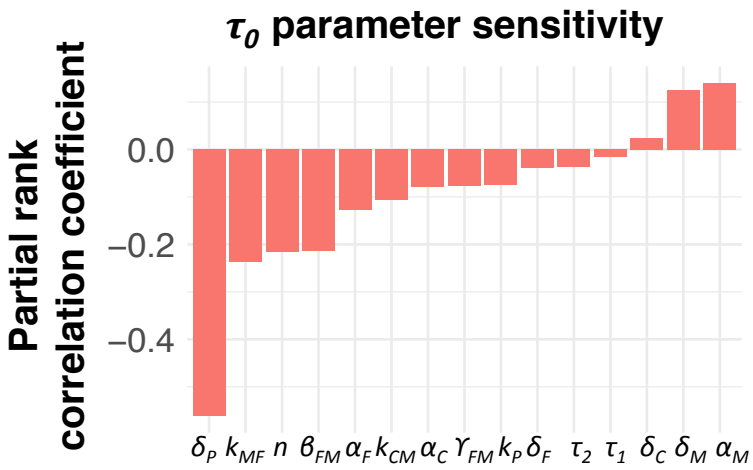

Figure 6.6: **Partial rank correlation coefficient for each parameter value considered correlated with the critical time for which a bifurcation occurs.** Correlation is taken partial to all other parameters, for the cases in which a bifurcation is predicted to occur. Parameter values considered were selected using Latin hypercube sampling, using 10^5 points in the parameter space, of which approximately 2300 had existence of a bifurcation.

6.2.5 A new mechanism for dynamically occurring oscillations

Through this analysis, the dependence of the parameter values on the critical time above which oscillatory behaviour emerges was determined. This dependence may be exploited by biological systems to generate oscillatory behaviour, as although the parameters governing the kinetics and delays present in a biological system are largely fixed, the rates of production and degradation can be changed dynamically [308, 351, 353, 354]. These changes may in turn change the critical time necessary for the system to exhibit oscillations, thereby moving the system into an oscillatory absorbing state, from a non-oscillatory state, and vice versa.

As an example of this, consider a time-varying value for α_C, where it is increased ten-fold from the baseline parameter values used in Figure 6.2, as may occur during particular developmental processes (e.g. those in which circRNA are hypothesised to function as miRNA sponges) [355]. In this case, the new system with a parameter value of $\alpha_C = 10$ has a critical time of $\tau_0 = 0.62$, which implies that the original system with $\tau_1 = \tau_2 = 0.5$ will oscillate in steady state. To visualise this change in absorbing state dynamically, consider the case where α_C is increased ten-fold only transiently between simulation times 50 and 150 min, and is 1 otherwise, as shown in Figure 6.7.

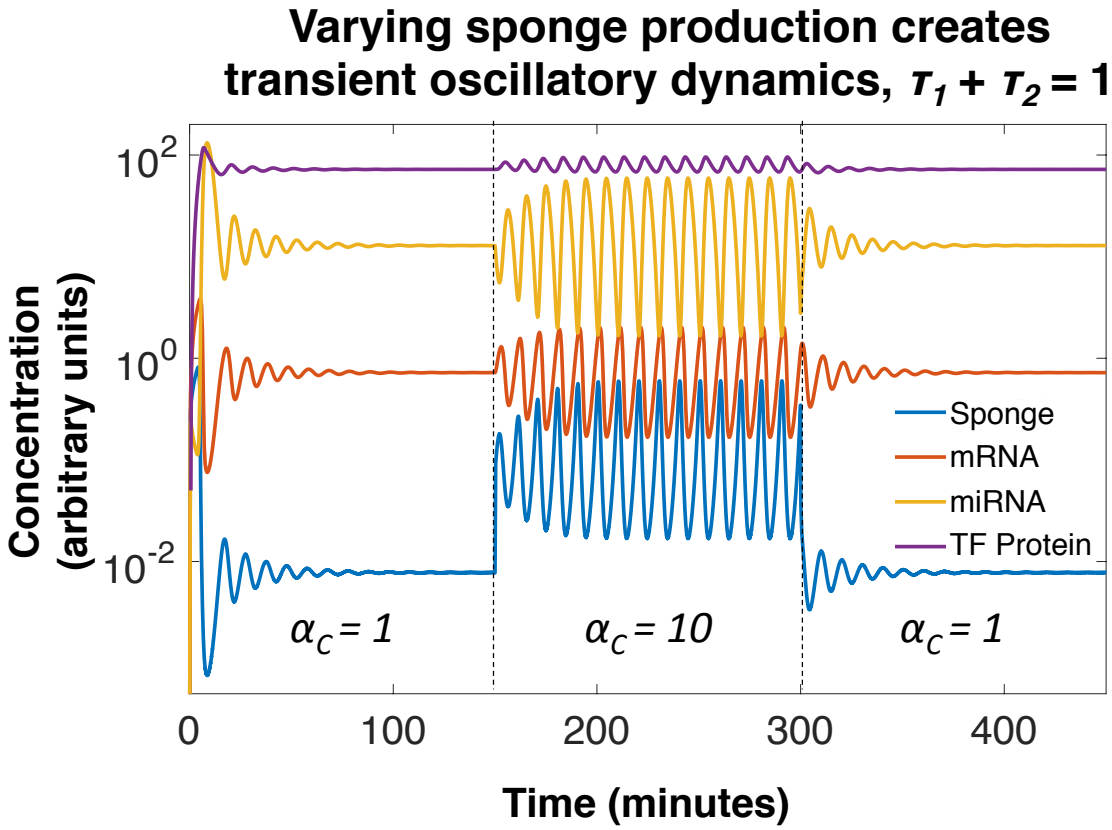

Figure 6.7: **A time-varying α_C generates transient oscillatory behaviour**. Here, a time varying value of α_C is used to illustrate the presence of a bifurcation. α_C is increased to 10 from an initial value of 1 between simulation time 150 and 300, between which oscillatory behaviour is the absorbing state, and is reduced to 1 otherwise, at which asymptotic stability predominates. Other parameter values are such that: $\delta_C = 0.01$, $\alpha_F = 1$, $\delta_F = 0.1$, $\alpha_M = 1$, $\delta_M = 1$, $k_P = 10$, $\delta_P = 0.1$, $k_{CM} = 10$, $k_{MF} = 0.1$, $\beta_{FM} = 200$, $\gamma_{FM} = 100$, and $n = 8$, with $\tau_1 = \tau_2 = 0.5$ as in Figure 6.2.

6.2.6 Stochastic simulation

All of the previous results shown rely on the assumption that there is a sufficient quantity of RNA molecules interacting in the system, such that a continuous approximation for their number is valid. However, the case where the number of molecules is small, and stochastic effects may predominate, must be considered as well, as it is possible that miRNA sponges and their targets are low in number within the cell. In this setting, the system is no longer well-described by the continuous variable ordinary differential equations as written in System 6.2, but rather is better-represented by a list of discrete events that occur at discretised time steps. The reaction 'events' and the associated rates at which they occur in the stochastic version of the system

are as described in System 6.2, with kinetic rate parameters on the right hand side, and a time delay indicated if present for that reaction. Each of the dynamic variables and parameters is as described above and in Table 6.2. The symbol $\emptyset$ on the left side of a reaction indicates *de novo* synthesis, and on the right side of a reaction indicates degradation.

$$\begin{array}{ll} \emptyset \rightarrow C : \alpha_C & M \rightarrow \emptyset : \delta_M M \\ C \rightarrow \emptyset : \delta_C C & \emptyset \rightarrow P : \alpha_P \\ \emptyset \rightarrow F : \alpha_F & P \rightarrow \emptyset : \delta_P P \\ F \rightarrow \emptyset : \delta_F F & C + M \rightarrow \emptyset : k_{CM} CM \\ \emptyset \rightarrow M : \alpha_M & M + F \rightarrow \emptyset : k_{MF} MF \\ \emptyset \rightarrow M : \alpha_{FM}(P) : \tau_1 & \emptyset \rightarrow P : k_P F : \tau_2 \end{array} \tag{6.46}$$

These events occur stochastically, and in general, systems described in this way have mean-field behaviour equivalent to their deterministic behaviour. Moreover, because of the presence of non-zero time delays τ_1 and τ_2, this system exhibits non-Markovian behaviour, and therefore the stochastic behaviour may not follow the mean-field approximation by the ODE system in the long-term. That is, there may be oscillatory behaviour in the stochastic case for a parameter regime where the deterministic model does not predict oscillations [356]. This phenomenon, of stochastic oscillations, is one of potential significance in the behaviour of these RNA networks, and is thought to contribute to the generation of circadian rhythms [357, 358], and in the Hes1 gene regulatory network [359].

To capture the potential for stochastic oscillations in the system, it was simulated numerically using the dde23 Runge-Kutta based solver in Matlab, noting that conventional analytic approaches to this problem are intractable as they require deriving and solving the Langevin equations derived from the reactions in System 6.46. To look for the presence of stochastic oscillations, one would then have to analytically derive the power spectra (based on the Fourier transform) for each of the stochastic time series C, F, M, and P, which would be intractable. Thus, numerical simulation was used going forward, using a modification of the Gillespie algorithm to simulate the system whilst accounting for time delays.

The Gillespie stochastic simulation algorithm enables the simulation of stochastic processes, by defining a list of potential reactions that may occur within the system, and the propensities at which these are likely to occur [360]. These propensities are dependent on the state of the system at the current time, and based on these, two random variables are drawn: the time interval until the next reaction, and the

next reaction that occurs. The state of the system is then updated, and the time advanced, and the process repeats itself. Because the state of the system at a future time is only dependent on the one previous time before it, the Gillespie stochastic simulation algorithm in its classical form, requires the system in question to have the Markov property. In time-delayed systems, such the system presented, the Markov property generally does not hold, as by definition of the time delay, the system will depend on previous states (those states specified by the time delay) in determining the future. As such, in order to simulate this system in the stochastic case, the modified Gillespie algorithm was considered, as has been used for similar purposes such as delayed mRNA gene networks and chemical reaction networks [358, 361].

The algorithm implemented, described in Algorithm 3, is based on the standard Gillespie algorithm, modified to handle the case of time-delayed reactions. In this algorithm, if a time-delayed reaction is chosen to occur based on the current state of the system, it is not executed until a future time, at which it is scheduled to occur (based on its time delay). This then necessitates the use of a queue system for future reactions, and if, based on the propensities of the system at the current time, the next instantaneous reaction is scheduled to occur after the next reaction in the queue, then the next reaction in the queue occurs instead and the time is updated to the time scheduled for this reaction to occur, at which point the algorithm restarts, and the previously selected instantaneous reaction does not occur. Using this algorithm, the stochastic behaviour of the system, with low numbers of each molecular species, was studied, for particular parameter sets.

To illustrate this, Figure 6.8 shows the mean field behaviour overlayed on $N = 100$ runs of the stochastic model, with one run of the stochastic model showing oscillatory behaviour. Furthermore, taking the Fourier transform and studying the power spectra through a periodogram of this stochastic signal reveals a strong subcomponent of an underlying oscillatory mode, whereas it was noted that the deterministic behaviour for this system did not show this oscillatory mode, as shown in Figure 6.9. The periodogram is a plot that displays the squared magnitude of the coefficient of the Fourier series at a given frequency on the vertical axis, plotted against this frequency on the horizontal axis. The periodogram for each of the 100 individual stochastic simulations, after subtracting the mean signal of each to standardise the baseline at 0 before taking the Fourier transform. Once the periodogram for each signal for each species in each of the stochastic runs was obtained, it was averaged at each frequency across each of these periodograms, and then normalised by a linear scaling factor, such that the maximum signal intensity was unity. These results, illustrated in Figure 6.9

capture the behaviour of each stochastic trajectory, and show that there is a high signal at the oscillatory mode corresponding to a period of 10-15 minutes, suggesting that stochastic oscillations are indeed a general phenomena for the system at these parameters, as seen in the individual trace in Figure 6.8. Thus, it is clear that for the system under consideration that such stochastic oscillations are a potential property, for realistic parameter values, and this adds a further dimension to the study for the potential behaviour of this system.

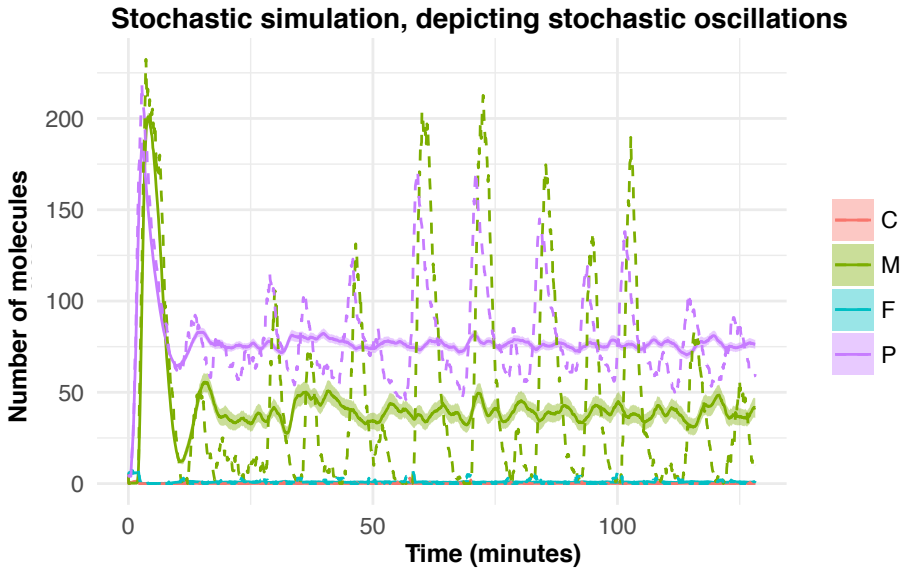

Figure 6.8: **Stochastic system dynamics, showing an individual trace of stochastic oscillations and mean field behaviour**. Averaged stochastic system dynamics do not show oscillations, but individual trajectories do. Dotted lines indicate an individual trajectory for a simulation, and bold lines are taken over an average of 100 runs, with standard error shaded around these lines. Parameter values used are the same as that of Figure 6.2, such that $\alpha_C = 1$, $\delta_C = 0.01$, $\alpha_F = 1$, $\delta_F = 0.1$, $\alpha_M = 1$, $\delta_M = 1$, $k_P = 10$, $\delta_P = 0.1$, $k_{CM} = 10$, $k_{MF} = 0.1$, $\beta_{FM} = 200$, $\gamma_{FM} = 100$, and $n = 8$, with $\tau_1 = \tau_2 = 0.5$, initial values chosen as 5 arbitrarily for all species.

6.3 Discussion

In this chapter, an existing reaction network was extended by the addition of a miRNA sponge. While the motif considered is theoretical in nature, each component has been shown to exist within the cell. The miRNA-mRNA-transcription factor negative feedback loop has been shown to occur with strong recurrence as a motif through global analyses of the human transcriptome [296]. Further, some of the strongest

Algorithm 3: Modified Gillespie stochastic simulation algorithm for time delayed dynamical system

```
function GillespieDelayed (reactions, params, initial conditions, t_end);
Input  : reactions, reaction parameters, initial conditions
Output: conc[], array of concentrations
        times[], array of times for which each concentration occurs
reaction_queue = empty queue
t = 0
conc[0] = initial conditions
while t < t_end do
    if t > next_reaction_time(reaction_queue) then
        reaction_occurring = next_reaction(reaction_queue)
        reaction_queue = reaction_queue.pop(reaction_occurring)
        t = next_reaction_time(reaction_queue)
        conc[t] = update_conc(conc[t-1],reaction_occurring)
    else
        propensities = f(reactions, conc)
        total_propensity = sum(propensities)
        rand_1 = random() %Random number between 0 and 1
        rand_2 = random() · total_propensity %Random number between 0
         and total propensity
        dT = -log(rand_1) / total_propensity %Time to next reaction

        %Select reaction from weighted probability
        reaction_occurring = g(propensities, rand_2)
        if isDelayed(reaction_occurring) then
            reaction_queue.push(reaction_occurring, scheduled_time)
        else
            t += dT %Time update
            conc[t] = update_conc(conc[t-1],reaction_occurring)
        end
    end
    return conc;
end
```

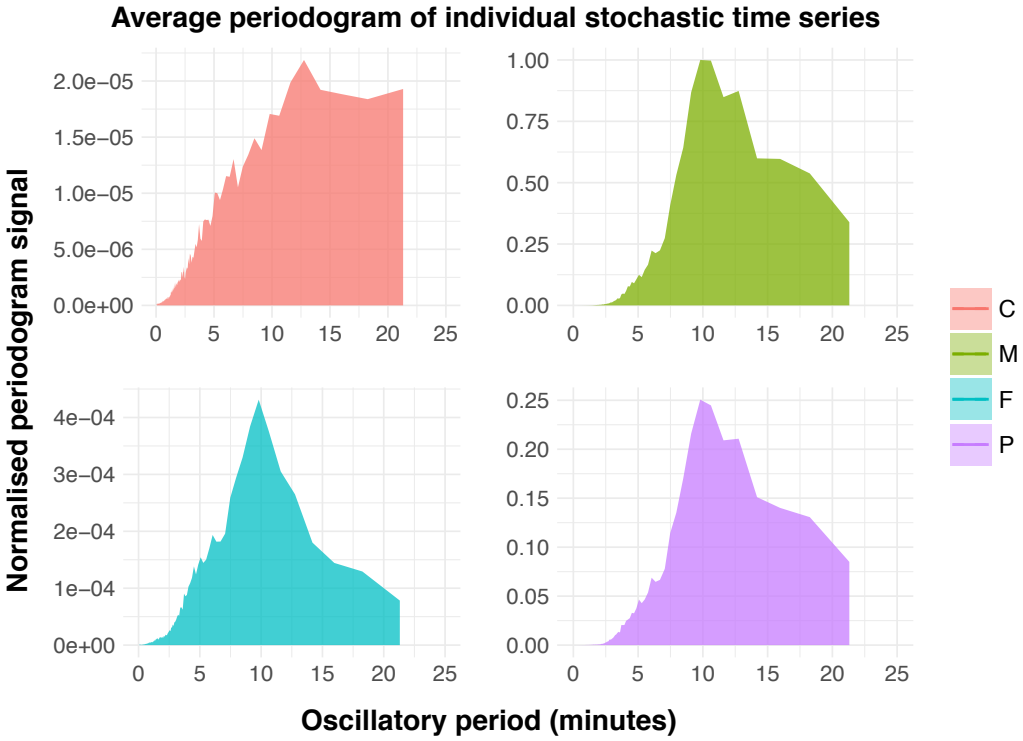

Figure 6.9: **Analysis of power spectra by periodogram highlights oscillatory behaviour.** Using the dynamics from stochastic simulations for miRNA sponge dynamics, the presence of underlying oscillatory modes, when the mean field behaviour predicts asymptotic stability was shown. Plots are of the average of 100 periodogram signal intensities, computed for each of the simulations of the stochastic model. Strong signal for an underlying oscillatory mode with period 10-15 minutes for the stochastic oscillations is evident, as corroborated by the individual series trace in Figure 6.8. As in Figure 6.8, parameter values used are: $\alpha_C = 1$, $\delta_C = 0.01$, $\alpha_F = 1$, $\delta_F = 0.1$, $\alpha_M = 1$, $\delta_M = 1$, $k_P = 10$, $\delta_P = 0.1$, $k_{CM} = 10$, $k_{MF} = 0.1$, $\beta_{FM} = 200$, $\gamma_{FM} = 100$, and $n = 8$, with $\tau_1 = \tau_2 = 0.5$.

experimental evidence for the function of particular circRNAs are as miRNA sponges, such as that of the circRNA cirRS-7 sponging miR-7 [247]. Combining these two lines of evidence, here the range of steady state behaviour, how it arises, and the effects of system parameters on these findings was studied. Thus, motivated by the findings of a circRNA (antisense to *HSP90AB1*) and its potential importance in regulating a switch in miRNA biogenesis in the previous chapter, it was considered how circRNA functioning as miRNA sponges may give rise to dynamic properties when incorporated into a feedback network motif.

6.3.1 Different species of miRNA sponges may confer different dynamical system properties

The findings from the mathematical model and its analysis are summarised in Figure 6.10, based on current knowledge of non-coding RNA that may function as RNA sponges. From this type of analysis, it becomes apparent that many possible dynamic behaviours are possible when considering the various types of miRNA sponges as components of gene networks. In particular, one of the emerging key features of circRNA that differentiates this species from the other types of ncRNA is their stability [362]. Because these RNA molecules do not have free ends, they are not subject to the same RNAse degrading enzymes present within the cell, and therefore can accumulate to a greater degree [363]. As shown, when functioning as miRNA sponges, these can act to destabilise the network of reactions, and push it closer to a state of oscillatory behaviour. This suggests a novel role of circRNA in processes involving development, where oscillations are both stage-dependent, and critical for organism genesis. Recent work involving circRNA characterisation has focused on their potential role in development, as knockdown of circRNA has been shown to have its strongest effects on neurogenesis of organisms, a process whereby oscillatory behaviour may be crucial [244, 251].

On the other hand, these results suggest that miRNA sponged by lncRNAs, which may have a short half-life within the cell (as identified through a recent genome-wide analysis of lncRNA half-lives by Clark et al.), are likely to exhibit greater stability and less propensity towards oscillatory behaviour, given their relatively higher degradation rate [364]. In effect, these lncRNA, if produced in targeted bursts, may provide tight temporal control of oscillatory behaviour, which may be crucial in regulating particular developmental processes, such as somitogenesis.

While experimental evidence substantiating these predicted effects for different parameter values on RNA levels does not yet exist, these network dynamics may be readily validated. For instance, a study functioning to validate the effects of changing a production rate could be performed *in vitro* by a synthetic gene construct under the control of two different promoters functioning at different strengths. Likewise, through the addition of degradation tags to the RNA molecules, there is the potential to assay the effects of different degradation rates. Further, through constructs showing different affinities for binding kinetics of miRNA sponge to miRNA, the effects of varying kinetic parameters such as the sponge binding rate can also be studied.

Such synthetic gene constructs have been used to study the predicted dynamics of oscillatory networks in *E. coli*, and validated the mathematical model predictions they were based upon [365].

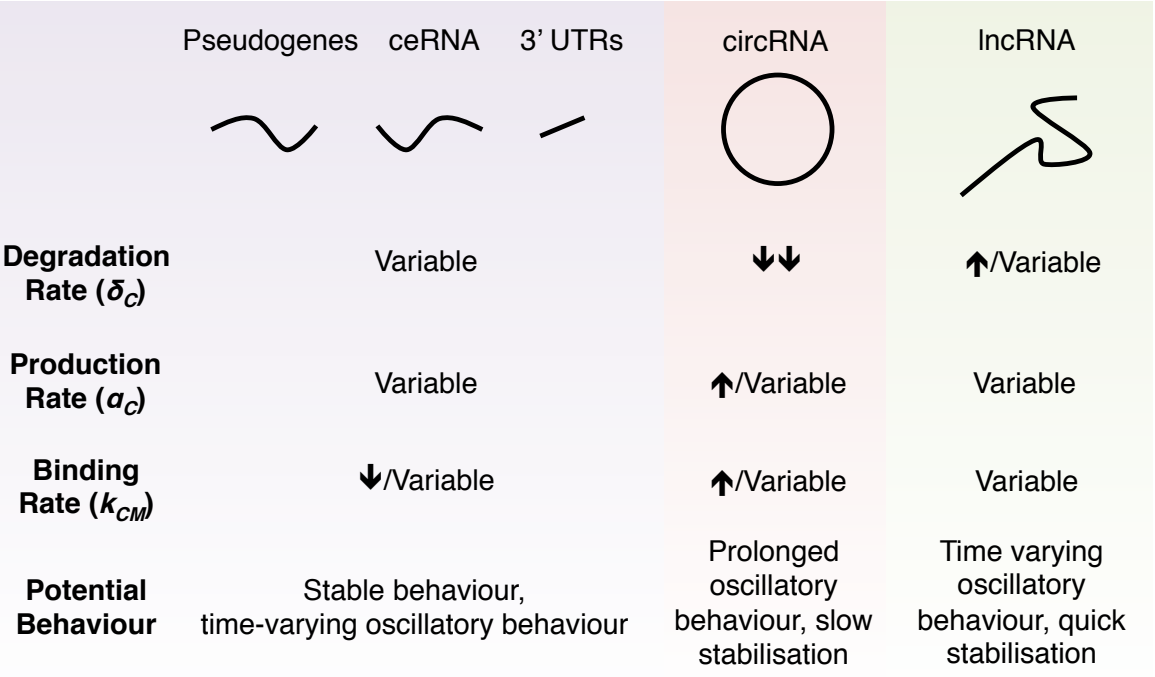

Figure 6.10: **Summary of potential behaviours for different ncRNA acting as miRNA sponges in reaction network**. Relationships between the dynamic parameters thought to occur for different ncRNA species functioning as miRNA sponges, and the effects of these parameter regimes has on system behaviour.

The above discussion is underscored by recent experiments showing both the existence and effect of specific miRNA sponges within human tissues in states of development and disease. As a result, further characterisation of the dynamic properties of these sponges is necessary to further elucidate the potential roles that they may play. As shown, these sponges may confer unique and important properties to the dynamic behaviour of RNA networks that have not yet been described. Particularly relevant to developmental biology, these results show how, for the first time, external control by a non-coding RNA species can give rise to periodically controlled oscillations. Previous models of oscillatory behaviour in RNA networks rely upon changing internal system parameters to achieve oscillatory behaviour, and have neglected the unique roles that a diversity of non-coding RNA can play in these dynamics. In addition, the networks on which these dynamics may occur have been characterised for a number of cancer-associated genes, and processes, such as the miR-200/*ZEB* module, which is known to be involved in the epithelial mesenchymal transition [299]. In

addition, a number of recent reports has shown the importance of circRNA in neural development, a process wherein oscillatory behaviour is known to occur [244, 251].

6.3.2 miRNA sponges in low copy number may be involved in the generation and maintenance of stochastic oscillations

As a result of the underlying biology of the interactions between the RNA, protein, and DNA species in this system, the mathematical description of this system involves delays.In practical terms, this means that the stochastic system dynamics can exhibit oscillatory behaviour, even in a parameter regime where the deterministic solution does not. This result, particularly for non-coding RNA, such as circRNA, which are thought to exist with low molecular concentrations within cells, suggests that oscillatory behaviour may be a common feature of these RNA species.

As a further theoretical exploration, the implications of extending the presented model to account for spatial differences in molecule count is considered. Biomolecules are heavily localised within the cell, and within this system as presented, dynamics arising from different spatial compartments has not been considered. For instance, because the majority of the molecular reactions occurring within this model are occurring in the nucleus, transport of the transcription factor protein into the nucleus must be accounted for. Furthermore, because of the effects of the reaction diffusion equations, system dynamics would differ between various sub-regions within the nucleus, such as within and outside of paraspeckles. Regions at the edges of the diffusive boundary would have smaller numbers of molecules, and therefore a greater propensity for stochastic oscillations. Thus, while a spatial model would give insight into the different behaviours that may arise, the dominant behaviours occurring within the spatial model would likely be the transition into the stochastic case from the deterministic case, as the number of molecules changes, based on the reaction-diffusion equations, resulting in disordered stochastic oscillations at a diffusive limit.

6.3.3 Implications for ncRNA-based therapeutics

The parameter sensitivity analysis in Figure 6.3 shows the key determinants of steady state levels for each of the species, and focussing on the values obtained for the miRNA sponge parameters, one can infer the impact of such a sponge on the steady states of the other species. This naturally leads to hypotheses regarding RNA therapeutic interventions. For example, in order to decrease the miRNA concentration, as opposed to increasing the binding kinetics of a miRNA sponge to the miRNA, the

model predicts that it would be more effective to increase the production rate of the sponge (or exogenously introduce a higher concentration of miRNA sponge). This encapsulates the robustness of such a feedback network in maintaining stable levels of a miRNA, and shows how disrupting such a network, if contributing to cancer, would really require a multi-pronged strategic approach, inhibiting multiple members of the feedback loop, as opposed to only a single member.

6.3.4 A novel experimental paradigm

The results shown provide fertile ground for generating hypotheses regarding the functional roles of the various miRNA sponge species. However, this was done within the confines of the limited evidence available at the current time for these species. In particular, characterisation of key kinetic parameters for these uncharacterised miRNA sponge species, through the generation of synthetic forms, could provide ample substrate for more clearly understanding their possible dynamics theoretically. More specifically, by first characterising the key rate parameters under stable microenvironmental conditions for miRNA production, binding kinetics, and degradation using fluorescent-labelled species and single-cell resolution imaging, a better sense of the rate constants involved could be obtained. Next, once these rate constants are obtained, using the results presented, one could determine whether a given system would have a propensity to oscillate or not with the introduction or inhibition of a miRNA sponge. Then, using synthetic approaches, such as an inducible promoter, for a miRNA sponge species, the sponge can be introduced or inhibited from the system, and different behaviours may be observed. Such behaviours would be more advantageously observed by a synthetic approach wherein the transcription factor activates a fluorescent protein, and such cells can then be monitored using time-lapse microscopy.

Further, because these miRNA sponges may lead to oscillatory behaviour, any experimental design implemented must be robust enough to capture this, which is why designs involving continuous monitoring such as time-lapse imaging are advantageous. Instead of supposing *a priori* that there will be asymptotically stable dynamics, multiple time points with a sufficiently fine resolution or continuous monitoring must be considered to determine whether these oscillations are present. If testing at multiple time points is considered, experimental guidance on the period of oscillations predicted should be used from the theoretical model, to ensure accurate sampling.

Overall, this chapter has shown how different miRNA sponges may result in different dynamical behaviours of a non-coding RNA network. Many RNA networks

of this form contain members innately involved in the progression of cancer and organismal development. This, together with emerging knowledge of the dysregulation of the non-coding transcriptome, suggests that there may be large differences in the dynamical behaviour of these RNA networks and their constituent species. This has potential implications in the development of anti-tumour therapies based on non-coding RNA, as an understanding of how the dynamic properties of ncRNA impact network behaviour, at least on a qualitative scale, will enable an understanding of what happens to behaviour when there is a perturbation of their levels by an external influence, such as a novel therapeutic.

6.4 Conclusions

Within this chapter, I have used mathematical modelling of a non-coding RNA-driven network. Using the tools afforded by modelling, I was able to elucidate the dynamic behaviour of this network in the case that a novel class of non-coding RNA, hypothesised to function as 'sponges' repressing miRNA, was active. Using this model framework, a set of mathematical relationships describing how this system evolved over time was defined, and through this, it was shown that a bifurcation exists in certain cases of parameter values. That is, for particular combinations of parameter values, the steady state behaviour of the network may switch from stable to oscillatory behaviour as rates of production, degradation, binding kinetics, and time delays vary. This holds implications for further evidence to the different potential roles for the various species of non-coding RNA that may function as miRNA sponges. I have characterised how the parameter values of the species involved in this network affect the steady state values, bifurcation existence, and critical time for oscillatory behaviours. As a result, I have shown how different miRNA sponges, with different parameters and time dependence of these parameters, may induce drastically different behaviours within the cell. Furthermore, when stochastic effects predominate, oscillatory behaviour for the molecular species involved may be more frequent than predicted by the deterministic analysis. This analysis therefore informs a potential key role for some species of non-coding RNA that function as miRNA sponges, particularly those with low degradation rate and high miRNA binding rate, such as circRNA.

This theoretical work has led to the generation of novel hypotheses for the potential functions of circRNA, which could be tested by circRNA in knockout experiments. On a theoretical note as well, because such miRNA networks and circRNA networks

have been found to occur independently in cancer, these results suggest how, in non-coding RNA networks involving genes crucial to the development and progression of cancer, by perturbation of a miRNA sponge, the dynamics may be altered greatly in steady state.

Chapter 7

Conclusions

The work presented in this thesis has, broadly, described the associative changes between different elements of the non-coding transcriptome in association with changes in the coding transcriptome. This was primarily accomplished through the re-analysis of existing large datasets with large sample sizes, allowing for high-dimensional multivariate statistical modelling. The information provided through this analysis of statistically associated changes may present readily testable hypotheses for non-coding RNA function in tumour biology.

7.1 Quality control is an important element of the gene signature validation process

The first portion of this thesis addressed a key issue related to gene signature validation; namely, the lack of quality control that is performed, when signatures are applied to independent validation cohorts. Because gene expression signatures are now being used as a part of clinical decision-making, it is crucial that there are quality control metrics to examine how well a signature captures an effect, before widespread adoption. Many of these issues relate back to the technicalities of how gene expression itself is measured; for instance there are a number of different technologies, with a plethora of computational preprocessing pipelines. To address this issue of gene signature reproducibility that arises as a result, *sigQC* was developed to encapsulate a set of statistical tests providing information of gene signature quality on a given dataset. To facilitate further widespread use, these statistical tests were wrapped into an R package, available through CRAN. This approach has enabled the quality control for many gene signatures on clinical datasets in a high-throughput manner, and

more recently, this tool has begun to be used as an adjunctive tool for gene signature refinement.

While the proposed protocol is not without limitations, it does provide an advance forward in beginning to provide standards to the field of gene signatures. The generality of the tool itself is a limitation, and the protocol itself has been designed for metagenes in particular, and only invokes relatively simple methods of gene signature scoring. Within the broader context of the field of gene signatures, *sigQC* does provide an easily-applicable set of metrics that can be widely applied, for at least a preliminary approach to gene signature quality control. Work by Berglund et al. initiated this discussion by providing similar metrics for principal component analysis-based signatures, [77]. *sigQC* extends this discussion to signatures that behave as metagenes, as well as further scoring metrics, however it is by no means comprehensive or sufficient to determine gene signature applicability. Thus, it is presented as a protocol that can be modified depending on the use of the signature under consideration, and one that will continue to evolve as time progresses.

7.2 A core set of miRNA associate statistically with cancer hallmark gene signatures

In the next chapter of this thesis, it was shown how mRNA gene expression signatures could be used to identify statistical associations with miRNA across tumour types. miRNA-gene signature associations across a set of 24 (quality-controlled) gene signatures representative of Hanahan and Weinberg's hallmarks of cancer [4, 5] were identified. Many of the miRNA identified as positively associated with the hallmarks of cancer had been previously validated as oncogenic, and associated with specific phenotypes, such as the hypoxia-associated miRNA. Next, the predicted mRNA targets of these miRNA were examined, and those that displayed consistent negative correlation with each miRNA were identified. These predicted targets with negative correlations across cancer types were shown to be enriched for tumour suppressor genes, suggesting that there is a potential miRNA-mediated mechanism by which tumour progression may occur, but any such assertion would require substantial experimental validation.

After having identified this level of regulation for a subset of 8 tumour suppressor genes, it was next examined whether there were cases for which the miRNA negatively associated with these tumour suppressor genes were still negatively associated with these genes' expression in the cases when mutation, methylation, and deletion of

these genes had not occurred. This revealed that the miRNA predicted to target these genes were upregulated in the cases across cancer types, where methylation was decreased, the genes were unmutated, and not deleted; particularly for *PTEN*, *FAT4*, and *CDK12*. Excitingly, these results provide the suggestion that cellular phenotype may in some cases be affected by miRNA levels, and that this possible effect may relate to tumourigenesis. However, as in the previous result, this cannot be concluded from the presented analysis, and further experimental validation is essential.

7.3 miRNA expression, maturation, and biogenesis show alterations in statistical association with hypoxia gene expression score

Motivated by the miRNA found to be highly altered in conjunction with hypoxia through the previous analysis of miRNA and the hallmarks of cancer, the statistically associated effects of these changes on the miRNA biogenesis pathway were examined. The work of Dr. Laura Winchester of the Buffa and Harris labs was extended, and a panel of 43 miRNA biogenesis genes was studied in relation to how they changed in copy number and expression in association with hypoxia gene expression signature score. Through this analysis, it was shown that hypoxia statistically significantly associated with an increase in *AGO2* copy number and expression and a corresponding decrease was observed in copy number and expression for *DDX5*, *DDX17*, and *TNRC6* across cancers, and specifically *DICER1* as well, in breast cancers. Following this analysis, the miRNA statistically associated with *AGO2* amplification, *DICER1* deletion, and hypoxia gene expression score increase, were identified. In particular, these were identified in three domains: one in which the mature miRNA overexpressed in association were identified, another where the mature:immature ratio of miRNA expression was examined, and a third domain where 5p to 3p ratio of miRNA expression was examined. It was shown that the mature miRNA expressed in association with *AGO2* amplification, *DICER1* deletion, and hypoxia gene signature score expression may show involvement with inflammation and oxidative phosphorylation, using the functional characterisation of miRNA from the prior chapter. Further, it was shown that for the miRNA showing a difference in arm selection preference, the arms chosen, preferentially negatively associated in correlation with tumour suppressor genes' expression, suggesting a possible link between their expression and tumour progression.

Overall this chapter focussed on the miRNA-level transcriptomic changes occurring in statistical association with hypoxia, and understanding how the changes to the biogenesis machinery in hypoxia co-associated with this. These changes were shown to associate statistically with factors that both may act to promote survival in hypoxia and tumour progression. However, this chapter is limited, in that the analysis is limited to the datasets considered, and that there may be inherent biases in the samples examined. In addition, these results, while suggestive of substantial results, provide correlative data only, and do not suggest any sort of causality. Indeed, this cannot be determined from the presented data and analysis alone, and for establishment of causality, experimental evidence is again required.

7.4 circRNA remain challenging to detect, but may show association with hypoxia gene expression signature

Continuing to characterise the roles of ncRNA in the hypoxic microenvironment, an existing dataset of TCGA breast cancers was re-examined, with RNA sequencing data remapped to circRNA, by Nair et al. [268]. This dataset was pre-processed using stringent filtering, and circRNA that had expression that was predictive of overall survival and hypoxia gene expression signature score were identified. One of these circRNA, a circRNA antisense to *HSP90AB1*, associated positively with *AGO2* as well, suggesting that it may have some mechanistic involvement with an alteration in biogenesis pathway, but this conclusion is speculative in the absence of experimental data.

The relationship between *AGO2* and *HSP90AB1* was further examined, and it was shown that largely, these two genes were positively correlated across cancer types in mRNA expression. Further, it was shown that this circRNA was strongly correlated to its sense transcript, suggesting the possibility of stabilisation by sense-antisense binding, but this conclusion would also require experimental evidence. Next, cell line data for the MCF-7 breast cancer cell line under normoxia and hypoxia was re-analysed to determine whether this circRNA could be detected, and to further understand the changes occurring to the circRNA transcriptome in hypoxia. It was shown that overall, circRNA increase in expression in hypoxia, but the method of RNA isolation has major effects on the quantity detected, and the circRNA themselves that are identified. In addition, a trend towards increasing levels of the circRNA antisense to *HSP90AB1* in hypoxia was shown, but this study was limited by small

sample size, precluding an analysis of statistical significance. In addition, the circRNA of interest, antisense to *HSP90AB1* was only identifiable from an analysis of polyA-selected RNA-seq data, the same as was available through the TCGA study. To summarise, the changes occurring to circRNA in hypoxia were characterised in for breast cancers, and it was shown that there is a potentially involved circRNA with the hypoxic response. These results were then associated with those of the previous chapter to examine the associative evidence between the expression of this circRNA and miRNA biogenesis genes, particularly *AGO2* amplification and *DICER1* deletion.

7.5 ncRNA network modelling reveals potential behaviours of miRNA sponges in different dynamic regimes

Having shown the behaviour of circRNA in hypoxia, the effects of miRNA sponging were next examined, as this represents a large class of interactions predicted for circRNA. A theoretical approach was taken, abstracting the concept to a generalised miRNA sponge, and considering the effect on network dynamics for a miRNA-mRNA-transcription factor motif overrepresented in the human transcriptome, and involving key genes associated with cancer. It was shown that the miRNA sponge in this network may act to cause transient oscillations, and in the stochastic regime, results in stochastic oscillations. Based on these findings, and a full examination of the possible dynamic behaviours of this network, it was hypothesised that the propensity towards oscillatory behaviour may lend important biological roles to certain species of miRNA sponges. In particular, the connections between miRNA sponges found experimentally, and their potential kinetic parameters, and the dynamics they might participate in, were presented. These results represent an advance in the mathematical modelling of non-coding RNA species, and suggest possible roles for key species of these molecules in development and cancer, but validation using particular experiments will be required to ascertain this.

7.6 Future directions

Through each of the sections presented within this thesis, a set of further questions, investigative paths, and potential projects has arisen. First, with respect to *sigQC*, further testing of the protocol on real-life use cases for a wide variety clinical and cell-line based datasets remains to be done. This level of large-scale intensive testing

would help to better inform the wider gene signatures community on the nuanced interpretation of the results of *sigQC*. Such work could define a guide for gene signature quality optimisation in different contexts, accounting for the technology and methodology by which a gene signature was derived, the dataset it is being tested on, and what it is meant to predict.

Next, with respect to the study of functionally characterising the miRNA statistically associated to the hallmarks of cancer, a number of miRNA-phenotype associations have been suggested by the methodology presented, and it remains to be determined whether definitive experimental evidence would support these hypotheses. In addition to this, the validation of miRNA showing possible exclusivity of negative statistical association for *PTEN*, *CDK12*, and *FAT4*-deficient tumours remains an open question, for which experimental validation is an attainable goal. Such validation would aim to show that the phenotype of the reduction in the tumour suppressor gene level could be reproduced by upregulating specific miRNA alone, in the absence of methylation, mutation, or gene deletion.

Next, with respect to the analysis of miRNA biogenesis genes and their alterations in hypoxia, and the role of circRNA in hypoxia, validation of predicted statistical associations is a potential future work. As stated, the work presented in this thesis hypothesises the possibility of a negative feedback loop, driven forward by hypoxia, leading to the upregulation of a circRNA antisense to *HSP90AB1*, stabilising the linear *HSP90AB1* mRNA transcript, enabling more HSP90AB1 protein to shuttle AGO2 to P-bodies. It was hypothesised that this increased activity of AGO2 and HSP90 in conjunction with hypoxia may act to mediated a change in miRNA biogenesis, towards one where AGO2 has taken over the slicing functions of DICER1 in the RISC, leading to preferential production of miRNA facilitating hypoxic adaptation and tumour progression. In the present work, these are suggestions and hypotheses as explicative of the statistical findings inferred, but these may be due to true signal or due to noise, and experimental validation could provide definitive proof. Moreover, repeating this analysis on larger and more varied datasets may provide further substantiation for these statistical associations and mechanistic hypotheses. These findings are being examined at present in a study currently being conducted by Dr. Shaunna Beedie of the Harris lab, performing quantitative PCR, Western blotting, and RNA-probe hybridisation experiments to analyse the changes in the levels of each of these key species in the hypoxic response.

The reanalysis of circRNA identified through a remapping of the TCGA breast cancer RNA sequencing data was limited by decreased circRNA counts, owing to

a polyadenylation selection step prior to the sequencing of these samples. For a more comprehensive characterisation of the circRNA involved in human cancers, it is clear that RNA-seq of polyA minus and ribo-minus samples is ultimately required. Therefore, obtaining these datasets, when available, with different computational pipelines is of great importance to further examine the potential role and statistical associations presented for circRNA.

Lastly, a mathematical model was presented for a non-coding RNA network, and this was used to study the dynamics of the network with the addition of a miRNA sponge. The miRNA sponge destabilised the fine-tuning effects of the miRNA within the network, and ultimately led to the development of oscillations within the system, though the existence of these depended on the dynamic parameters of the miRNA sponge. Thus, one aspect of future work is to study these systems experimentally, by constructing these gene networks synthetically, and analysing how changing the dynamic properties of the miRNA sponge changes overall behaviour. Further, this analysis was for a single overrepresented motif, and there are many others that are conserved throughout the transcriptome. Studying the possible effects of miRNA sponges on these networks is also a crucial element of future work that needs to be done, as it may be that their addition alters dynamics in these systems to an even greater degree.

7.7 Concluding remarks

The computational approach presented in this thesis carried within it new methods and techniques for establishing statistical association between the phenotypic hallmarks of cancer, and the genotypes that underlie these. Clinical tumour specimens present a huge promise to advancing personalised medicine, but great care needs to be taken when analysing these datasets. Issues of reproducibility, validation, and technical considerations need to be considered very critically and accounted for, before any firm conclusions from a data-driven approach can be made, though using large sample sizes can help to separate the signal from the noise. It is my hope that through this thesis I have provided statistical evidence that may represent small scientific step forward.

Appendix A

Appendix: sigQC

A.1 Materials

A.1.1 Equipment

Hardware:

- Personal computer, capable of running R version 3.3.0 or higher

Software:

- R version $\geq$ 3.3.0, available to install from https://www.r-project.org/
- *sigQC* package, available to download from https://cran.r-project.org/web/packages/sigQC/index.html

A.1.2 Equipment setup

R software installation:

- Download and install the latest version of R from https://www.r-project.org/, or the freely available RStudio from https://www.rstudio.com/.

***sigQC* installation:**

- To install the *sigQC* package, execute the following command in R or RStudio:

```
install.packages("sigQC")
```

Input formats and usage:

The primary user-accessible function of *sigQC*, make_all_plots, expects a number of inputs, the format of each of which is defined in Table A.1 as well as the package documentation. Further, once installed and with all data loaded into the appropriate variables, use of the package is accomplished with the following commands in R or RStudio:

```
library("sigQC")
make_all_plots(gene_sigs_list, mRNA_expr_matrix, names_sigs,
    names_datasets, covariates, thresholds, out_dir, showResults, origin,
    doNegativeControl, numResampling)
```

Downloading of sample data and code:

Sample randomly generated data and code can be found in the package vignette example that is available for download with the package at
https://cran.r-project.org/web/packages/sigQC/index.html

A.2 Procedure

A.2.1 Preparation of input data:

- The input data should consist of lists of expression matrices of at least 2 samples each, and should be pre-normalised, log-transformed (as per gene signature requirements), and standardised if required. Care should be taken to ensure that genes of interest are present in the dataset and not reported primarily as NA values. Batch effects present across multiple datasets compared are not required to be corrected by the user, as *sigQC* is designed to test on multiple independent datasets. However, if a single dataset with subcomponents affected by batch effects is included, this should be corrected internally before use.

- Additionally, the signatures to be tested must be annotated in a manner consistent with the input data. Furthermore, any specific expression thresholds for expression (other than global median) should be computed, as this is the default the package uses as an expression cutoff.

- Lastly, any additional annotation data to be used alongside the expression heatmaps should be identified, and loaded into the appropriate matrices with color descriptors as specified in the package documentation.

Variable Name	Default value	Description
gene_sigs_list	None	A list of gene signature matrices, representing the gene signatures to be tested.
mRNA_expr_matrix	None	A list of expression matrices, one for each dataset
names_sigs	NULL	The names of the gene signatures (e.g. Hypoxia, Invasiveness), one name per each signature in gene_sigs_list.
names_datasets	NULL	The names of the different datasets contained in mRNA_expr_matrix
covariates	NULL	A list containing a sub-list of 'annotations' and 'colors' which contains the annotation matrix for the given dataset and the associated colours with which to plot in the expression heatmap.
thresholds	NULL	A list of thresholds to be considered for each data set, default is median of the data set. A gene is considered expressed if above the threshold, non-expressed otherwise. One threshold per dataset, in the same order as the dataset list.
out_dir	tempdir()	A path to the directory where the resulting output files are written
showResults	TRUE	Tells if open dialog boxes showing the computed results. Default is TRUE
origin	NULL	Tells if datasets have come from different labs/experiments/machines. Is a vector of characters, with same character representing same origin. Default is assumption that all datasets come from the same source.
doNegativeControl	TRUE	Logical, tells the function if negative and permutation controls must be computed.
numResampling	50	Integer for the number of re-samplings while computing negative and permutation controls

Table A.1: **Description of input variables to *sigQC* function make_all_plots().**

A.2.2 Creation of input variables:

- Preparation of gene signatures list: The gene signatures considered should each be $k \times 1$ sized character matrices for a signature of length k genes. The in-

dividual elements of these matrices should be the gene names (and should be consistent with the naming convention for genes as named in the row names of the expression matrices). These matrices should all then be saved into a single R list variable, such that each matrix is one element of the list of gene signatures, and named to describe the gene signature contained in this matrix.

- Preparation of gene expression list: The gene expression matrices considered should be matrices with rows as the genes and columns as the individual samples. As described above, the expression matrices should be normalised, batch-corrected, log-transformed, and standardised as needed, prior to use of *sigQC*. The row names of these matrices should be the gene names, and naming conventions consistent with those allowable with R matrix naming requisites are permitted. These are the same gene names that will be displayed on the produced plots. Each gene expression matrix for each dataset considered should be saved as an element of a single list variable in R, with each element of this list set as one of the gene expression matrices, and named to describe this dataset.

- If alternative names (other than those used for list indexing) are desired for the plots produced by *sigQC*, there is the option to set the *names_sigs* and *names_datasets* variables, which are vectors containing the desired names of the signatures and datasets (ordered in the same way as the list variables they represent).

- During the plotting of heatmaps showing the expression of the signature genes across samples in the various datasets, if it is desired to have annotation rows at the top of the heatmaps, indicating sample characteristics, this is possible through the *covariates* input variable. This is a list with one sub-list element per dataset. Each of these sub-lists contain two matrix elements named 'annotations' and 'colours,' describing the annotation values for each of the samples and the associated colours to be used in the plotting. For further information about this variable, the user is referred to the documentation for the ComplexHeatmap R package, as this is the same *covariates* variable as used in this package.

- In the exploration of gene expression, if expression above a particular threshold is required (e.g. a noise value), the thresholds variable can be set (with threshold values for each dataset in the same order as they appear in the list of datasets).

If this value is not set, it is defaulted to the median expression of all genes across all samples in each dataset.

- The output directory variable should be set as a string for a file path that is reachable from the current directory in the *out_ dir* variable. If no value is set, this defaults to the temporary directory given by R in tempdir().

- If the user wishes to see results in the R graphics windows as they are created, the *showResults* parameter can be set to TRUE (default), otherwise it can be set to FALSE.

- If the datasets have been derived from different labs and experimental setups, the origins parameter can be set to indicate this. This is only used in the computation of the rank product statistic when comparing autocorrelation of genes across datasets, to identify consistently poorly correlated signature genes, to account for batch effects between datasets. It should be set as a vector of numbers or characters, with each element indicating numerically the origin of a dataset, in the same order as they appear in the dataset list input variable.

- For a comparison to the null distributions via bootstrap resampling for random sets of genes and permutations of the gene signature labels, the *doNegativeControl* variable should be set to TRUE, otherwise it should be set to FALSE. The *numResampling* variable is set to the number of bootstrap resampling runs to be done, if the *doNegativeControl* variable is true.

A.2.3 Running of *sigQC* package:

- With the input data pre-processed and in the appropriate variables, the principal function of the *sigQC* package can be run, with the following command:

```
library("sigQC")
make_all_plots(gene_sigs_list, mRNA_expr_matrix, names_sigs,
    names_datasets, covariates, thresholds, out_dir, showResults)
```

- This produces, in the output directory or graphically displayed directly to the user if desired, a number of plots in PDF files which may be analysed as described in the subsequent steps. The package also creates an output file 'log.log' in the output directory, a text file, which summarises the run, and reports any errors that may have occurred if they are not printed to the console. This should

be consulted if any issues are encountered in the running of this principal function and for troubleshooting purposes.

A.3 Timing

The timing of *sigQC* functions varies, depending on the number of datasets and signatures analysed, from few minutes (for the examples shown here) to hours (for concomitant analysis of several datasets and signatures, and high number of replicate resampling).

A.4 Troubleshooting

A.4.1 Installation:

Issues may be experienced if the ImageMagick dependency is not installed on the user's system (particularly for Windows systems). To install this dependency, please follow directions at:

http://imagemagick.org/script/download.php.

A.4.2 Step 3:

Issues may be experienced with input data not conforming to the format required by *sigQC*. If this occurs, the package will alert the user with an error message describing the nature of the discrepancy. For example, common errors may include the following:

- Gene signatures must be formatted as a list of matrices, of dimension k rows by 1 column, for a signature of length k genes. Inputting a single list as a vector will cause an error to the program.
- Datasets must also be formatted as lists of matrices, such that genes are the rownames of the dataset, and samples are organised by columns of the dataset.
- Gene signatures and datasets must be annotated in the same way, as if the names of the genes of a signature are not found in a dataset, the computation will not continue.
- Care must be taken to ensure that NA valued genes are removed as optimally as possible, as if there are too many values in the expression matrix for the gene signature are NA, calculations dependent upon singular value decomposition (e.g. principal components analysis) cannot be carried out.

A.5 *sigQC* availability

The *sigQC* package has been made available for download from CRAN at https://cran.r-project.org/package=sigQC, and can be cited through this publication.

A.6 Pseudocode for radar plot metrics

Define an m-dimensional array, $e = [e_1, \ldots, e_m]$ as the gene expression data relative to a single sample, such that e_k is the expression value of gene k in the given sample. In this way, the full dataset may be defined as the bi-dimensional matrix$E = [e_1, ..., e_n]$, where n is the number of samples and e_{ij} is the expression value of gene i in the j-th sample. Similarly, let $E = [e_1, ..., e_m]^t$ the same matrix, where e_k is an n-dimensional array containing the expression data of a single gene across all n samples and $(.)^t$ indicates the transpose of a matrix. Finally, denote by $R = [r_1, ..., r_n]$ the reduced gene expression matrix containing only the expression of the genes included in the assessed signature so that $r_k = [r_1, \ldots, r_l]$, where $l \leq m$.

A.6.1 Ratio of Med. SD

1. Compute the standard deviation (σ_1) of each signature gene across all samples
2. Denote by α the median of the standard deviations
3. For every gene, compute the standard deviation (σ_2) across all samples
4. Denote by β the median of the standard deviations
5. Return the absolute value of $\alpha/(\alpha + \beta)$

A.6.1.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
$\sigma_1 = [\sigma_{11}, ..., \sigma_{1l}]$ = l-dim array
$\sigma_2 = [\sigma_{21}, ..., \sigma_{2n}]$ = m-dim array
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  | $\sigma_1(i)$ = standard deviation(r_i)
}
for ($j = 1;\ j \leq m;\ j = j + 1$) {
  | $\sigma_2(j)$ = standard deviation(e_j)
}
α = median(σ_1)
β = median(σ_2)
return $|\alpha/(\alpha + \beta)|$

A.6.2 Med., Z-Med. Score Cor.

1. Compute the median of each signature gene across all samples
2. Normalise the input matrix using the z score
3. Compute the median of each signature gene in the normalised matrix across all samples
4. Compute the Spearman correlation between the 2 median arrays
5. Return the absolute value of the Spearman correlation coefficient

A.6.2.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
med, med_z, μ, σ = l-dim arrays
$Z = [z_1, ..., z_n] = [z_1, ..., z_l]^t$ = normalised matrix
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  $\text{med}(i) = \text{median}(r_i)$
  $\mu(i) = \text{mean}(r_i)$
  $\sigma(i) = \text{standard deviation}(r_i)$
}
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  for ($j = 1;\ j \leq n;\ j = j + 1$) {
    $Z(i, j) = (r_{ij} - \mu(i))/\sigma(i)$
  }
}
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  $\text{med}_z(i) = \text{median}(z_i)$
}
$\rho = \text{correlation}(\text{med}, \text{med}_z)$
return $|\rho|$

A.6.3 Mean, PCA1 Score Cor.

1. Compute the mean of each signature gene across all samples
2. Compute the first principal component (PCA1) of each signature gene across all samples
3. Compute the Spearman correlation between the mean and PCA1 arrays
4. Return the absolute value of the Spearman correlation coefficient

A.6.3.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
μ, pca1 = l-dim arrays
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  $\mu(i) = \text{mean}(r_i)$
  $\text{pca1}(i) = \text{first principal component}(r_i)$
}
$\rho = \text{correlation}(\mu, \text{pca1})$
return $|\rho|$

A.6.4 PCA1, Z-Med. Score Cor.

1. Compute the first principal component (PCA1) of each gene across all samples
2. Normalise the input matrix using the z score
3. Compute the median for each signature gene across all samples in the normalised matrix
4. Compute the Spearman correlation between the PCA1 and median arrays
5. Return the absolute value of the Spearman correlation coefficient

A.6.4.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
pca1, med_z, μ, σ = l-dim arrays
$Z = [z_1, ..., z_n] = [z_1, ..., z_l]^t$ = normalised matrix
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  pca1(i) = first principal component(r_i)
  $\mu(i)$ = mean(r_i)
  $\sigma(i)$ = standard deviation(r_i)
}
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  for ($j = 1;\ j \leq n;\ j = j + 1$) {
    $Z(i, j) = (r_{ij} - \mu(i))/\sigma(i)$
  }
}
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  $\text{med}_z\ (i)$ = median(z_i)
}
ρ = correlation(pca1, med_z)
return $|\rho|$

A.6.5 Mean, Med. Score Cor.

1. Compute the mean of the signature genes for each sample
2. Compute the median of the signature genes for each sample
3. Compute the Spearman correlation of the mean and median arrays
4. Return the absolute value of the Spearman correlation coefficient

A.6.5.1 Pseudocode

$R = [r_1, ..., r_n]$
μ, med = n-dim arrays
for ($j = 1;\ j \leq n;\ j = j + 1$) {
  μ =mean(r_j)
  med = median(r_j)
}
ρ =correlation(μ, med)
return $|\rho|$

A.6.6 Med. Autocor.

1. Compute the autocorrelation of the reduced gene expression matrix
2. Return the absolute value of median of all correlations coefficients

A.6.6.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
$A = l \times l$ matrix
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  for ($j = 1;\ j \leq l;\ j = j + 1$) {
    $A(i, j)$ =correlation(r_i, r_j)
  }
}
return |median(A)|

A.6.7 Med. Prop. Expressed

1. Compute the median of the dataset
2. For each gene, check if expression is greater than median
3. For each gene, count the proportion over all samples
4. Return the median over the array of proportions

A.6.7.1 Pseudocode

$E = [e_1, ..., e_n], R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
prop = m-dim array
med = median(E)
$C = l \times n$ zeros matrix
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  for ($j = 1;\ j \leq n;\ j = j + 1$) {
    if ($r_{ij} > med$) **then**
      $C(i, j) = 1$
  }
}
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  prop(i) = count($C(i)$)/n
}
return median(prop)

A.6.8 Med. non-NA Prop

1. Count the number of times each gene in the signature is expressed over all samples
2. For each gene, compute the expression proportion over all samples
3. Return the median over the array of proportions

A.6.8.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
prop = m-dim array
$C = l \times n$ zeros matrix
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  for ($j = 1;\ j \leq n;\ j = j + 1$) {
    if $(r_{ij} \neq NA)$ **then**
      $C(i, j) = 1$
  }
}
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  prop(i) = count($C(i)$)/n
}
return median(prop)

A.6.9 Coef. of Var. Ratio

1. Compute the standard deviation (σ) for each signature gene across all samples
2. Compute the mean (μ) for each gene across all samples
3. Compute the coefficient of variation ($c_{v1} = \sigma/\mu$) for each signature gene across all samples
4. Denote by α the median of the coefficients of variation
5. For each gene, compute the coefficient of variation (c_{v2}) across all signature genes
6. Denote by β the median of all c_{v2}
7. Return the absolute value of $\alpha/(\alpha + \beta)$

A.6.9.1 Pseudocode

$E = [e_1, ..., e_n] = [e_1, ..., e_m]^t, R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
c_{v1} = l-dim arrays
c_{v2} = n-dim arrays
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  | $c_{v1}(i)$ = standard deviation(r_i)/ mean(r_i)
}
α = median(c_{v1})
for ($j = 1;\ j \leq m;\ j = j + 1$) {
  | $c_{v2}(j)$ = standard deviation(e_j)/ mean(e_j)
}
β = median(c_{v2})
return $|\alpha/(\alpha + \beta)|$

A.6.10 Prop in top 50% var.

1. Compute the standard deviation (σ) for each gene across all samples
2. Compute the mean (μ) for each gene across all samples
3. Compute the coefficient of variation ($c_v = \sigma/\mu$) for each gene across all samples
4. Rank the c_v
5. Return the proportion of signature genes with c_v in the top 50% of the rank

A.6.10.1 Pseudocode

$E = [e_1, ..., e_n] = [e_1, ..., e_m]^t$
c_v = m-dim array
c = l-dim zero array
for ($i = 1;\ i \leq m;\ i = i + 1$) {
  $c_v(i)$ = standard deviation(e_i)/ mean(e_i)
}
q = quantile$_{0.5}$ (c_v)
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  if ($c_v(i) \geq q$) **then**
    $c(i) = 1$
}
return count(c)/l

A.6.11 Prop in top 25% var.

1. Compute the standard deviation (σ) for each gene across all samples
2. Compute the mean (μ) for each gene across all samples
3. Compute the coefficient of variation ($c_v = \sigma/\mu$) for each gene across all samples
4. Rank the c_v
5. Return the proportion of signature genes with c_v in the top 25% of the rank

A.6.11.1 Pseudocode

$E = [e_1, ..., e_n] = [e_1, ..., e_m]^t$
c_v = m-dim array
$c = l$-dim zero array
for ($i = 1;\ i \leq m;\ i = i + 1$) {
  $c_v(i)$ = standard deviation(e_i)/ mean(e_i)
}
q = quantile$_{0.75}$ (c_v)
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  if ($c_v(i) \geq q$) **then**
    $c(i) = 1$
}
return count(c)/l

A.6.12 Prop in top 10% var.

1. Compute the standard deviation (σ) for each gene across all samples
2. Compute the mean (μ) for each gene across all samples
3. Compute the coefficient of variation ($c_v = \sigma/\mu$) for each gene across all samples
4. Rank the c_v
5. Return the proportion of signature genes with c_v in the top 10% of the rank

A.6.12.1 Pseudocode

$E = [e_1, ..., e_n] = [e_1, ..., e_m]^t$
c_v = m-dim array
$c = l$-dim zero array
for ($i = 1;\ i \leq m;\ i = i + 1$) {
  | $c_v(i)$ = standard deviation(e_i)/ mean(e_i)
}
q = quantile$_{0.90}$ (c_v)
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  | **if** ($c_v(i) \geq q$) **then**
  | | $c(i) = 1$
}
return count(c)/l

A.6.13 Skew Ratio

1. Compute the skewness (α) of mean of each signature gene, across all samples
2. Compute the skewness (β) of mean of each gene, across all samples
3. Return $|\alpha|/(|\alpha| + |\beta|)$

A.6.13.1 Pseudocode

$E = [e_1, ..., e_n] = [e_1, ..., e_m]^t, R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
μ = m-dim array
for ($i = 1;\ i \leq m;\ i = i + 1$) {
  | $\mu(i)$ = mean(e_i)
}
α = skewness[$\mu(r_j)$]
β = skewness[$\mu(e_j)$]
return $|\alpha|/(|\alpha| + |\beta|)$

A.6.14 Prop Var by PCA1

1. Compute the principal component of every signature gene across all samples
2. Return proportion of variance explained by first principal component

A.6.14.1 Pseudocode

$R = [r_1, ..., r_n] = [r_1, ..., r_l]^t$
pca1= l-dim arrays
for ($i = 1;\ i \leq l;\ i = i + 1$) {
  | pca1(i) = first principal component(r_i)
}
return variance_prop(pca1)

Appendix B

Appendix: miRNA hallmarks

B.1 Listing of genes included in each gene signature, COSMIC tumour suppressor genes, and oncogenes

The listing for the Entrez IDs used for all gene signatures considered are provided in the text files contained within the gene_signatures subfolder of the supplementary .zip file, available for download at `https://github.com/andrewdhawan/miRNA_hallmarks_of_cancer/`. The lists of COSMIC tumour suppressor genes and oncogenes may be found in text files wtihin the COSMIC subfolder within the supplementary .zip file.

B.2 *sigQC* Gene signature quality control summary plots

Here, radar plots summarising the various gene signature quality control metrics implemented by the *sigQC* R package are presented. In general, signature quality is reflected by the overall closeness to the outer rim of the radar plot, for each of the 14 metrics considered. Figures B.1- B.9 contain signature quality control plots grouped by approximate biological categories: angiogenesis, apoptosis, energetics, genome instability, growth suppressors, immortality, inflammation, invasion, and proliferation.

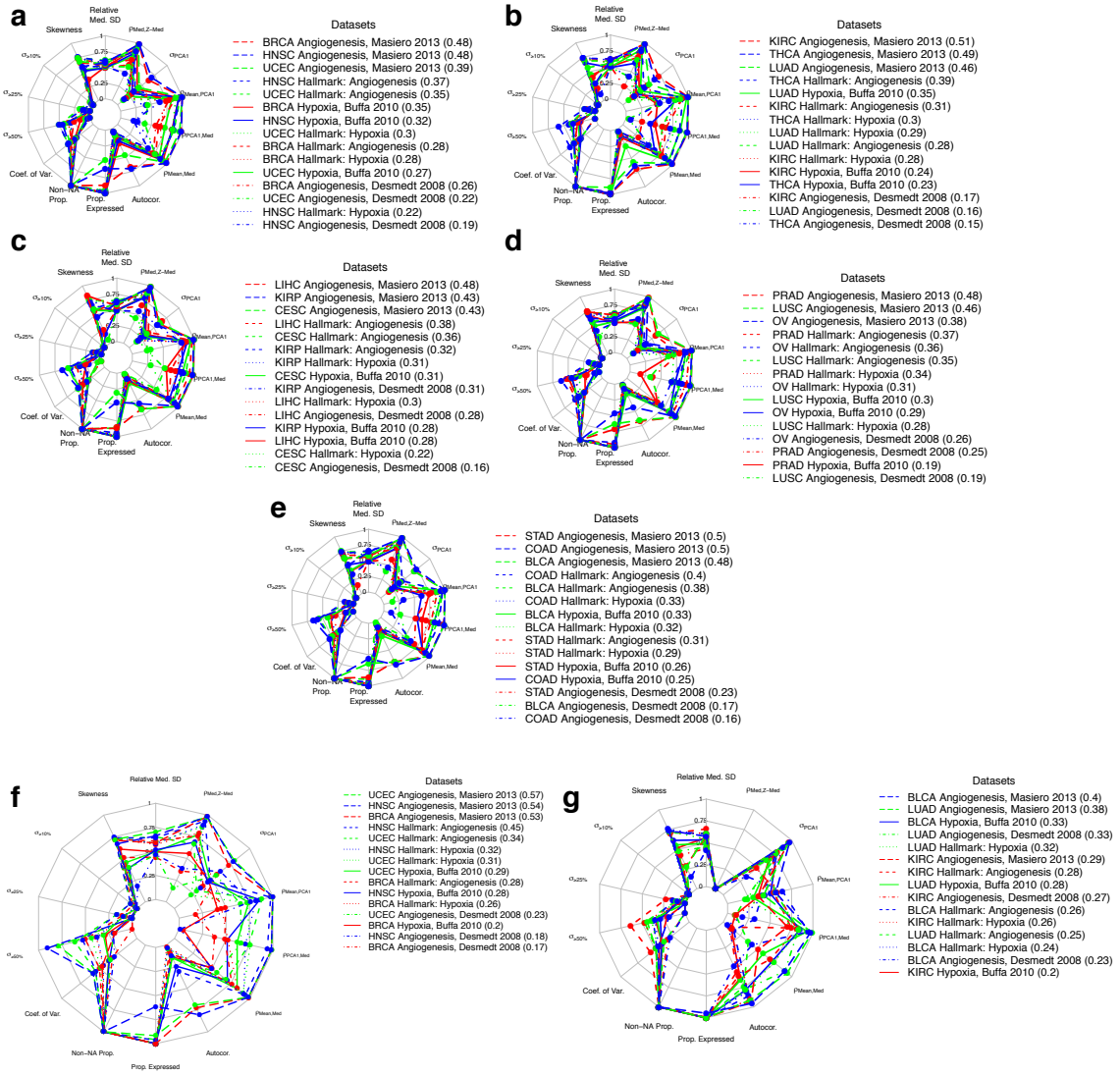

Figure B.1: ***sigQC* radar plots for angiogenesis-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

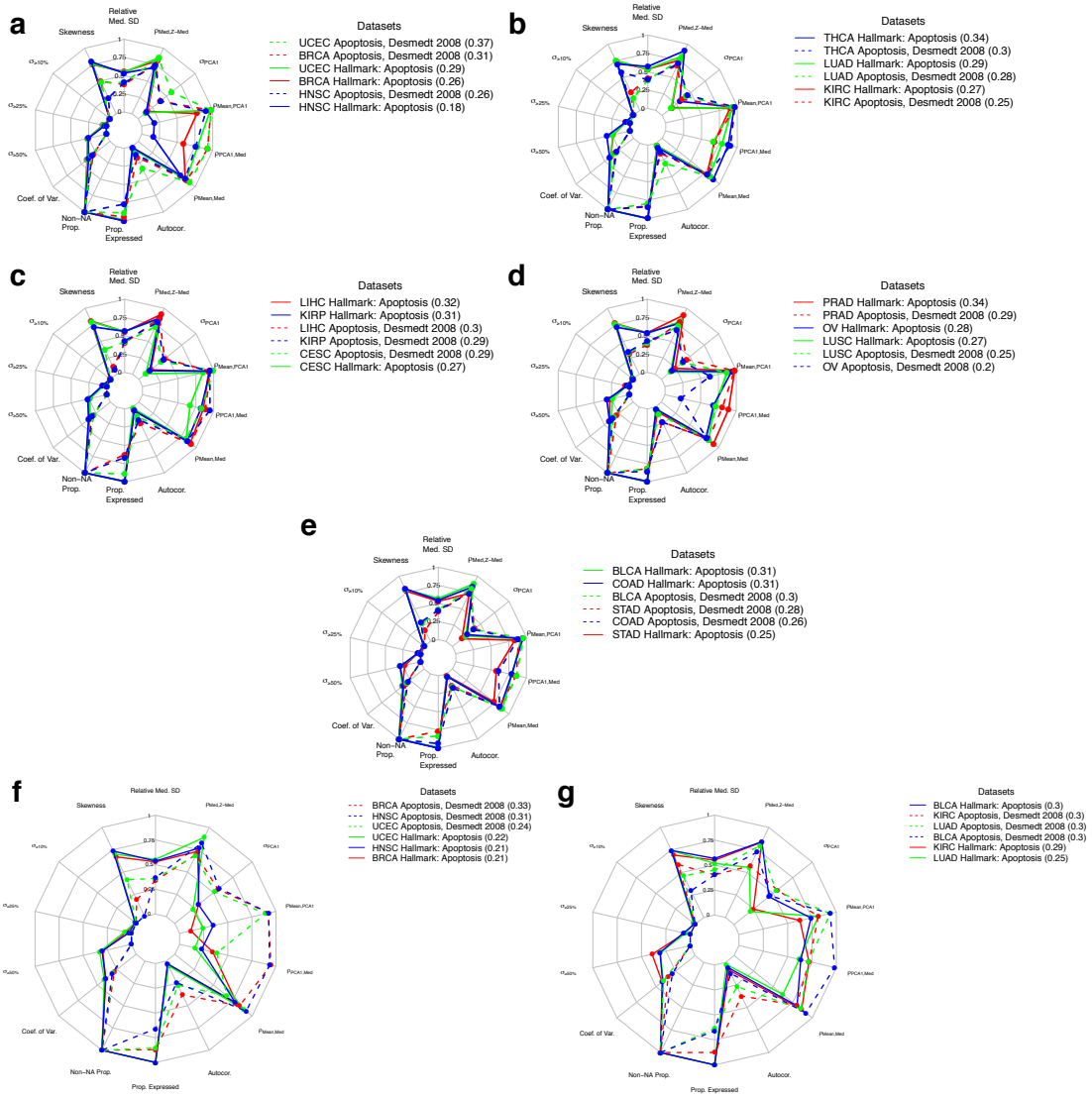

Figure B.2: ***sigQC* radar plots for apoptosis-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

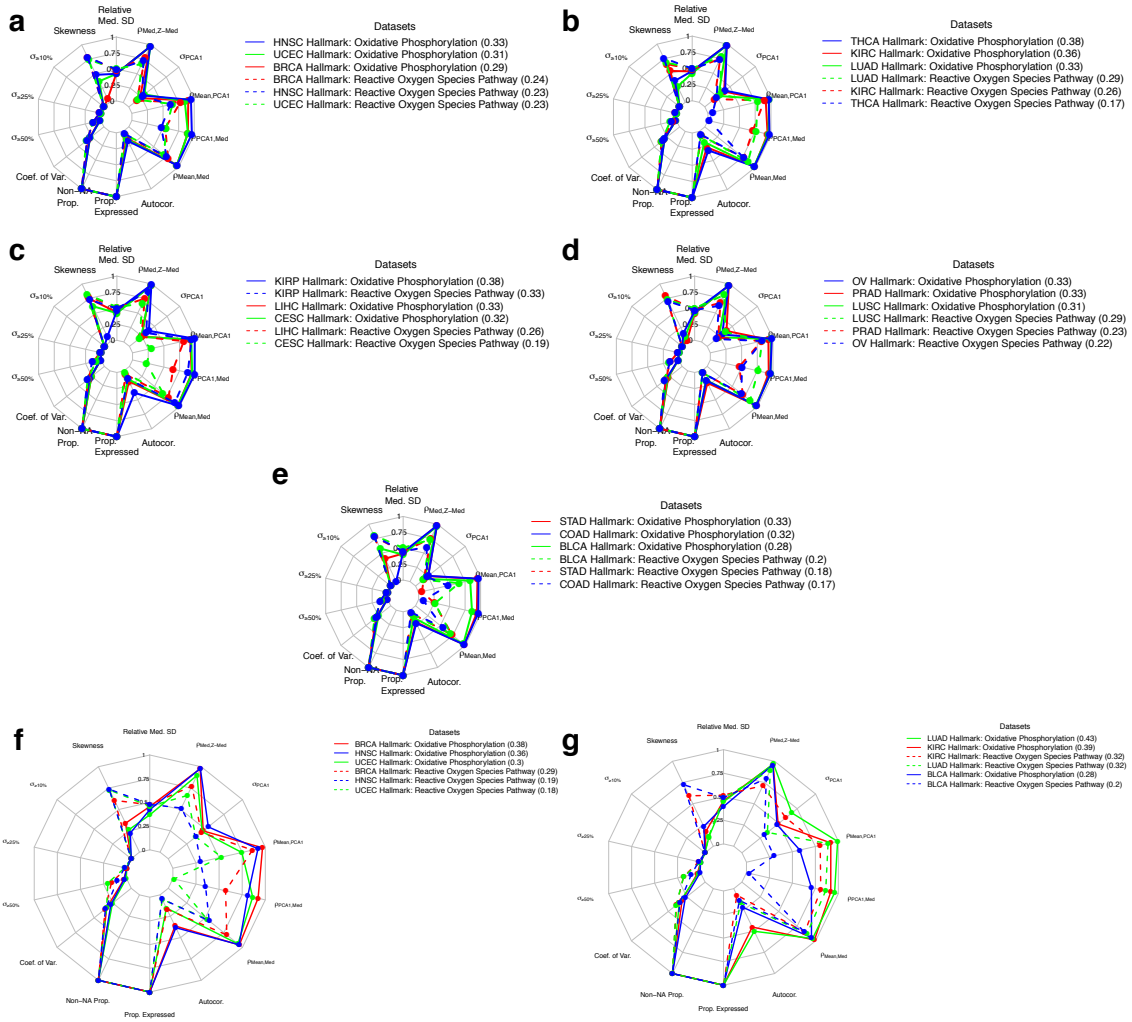

Figure B.3: ***sigQC* radar plots for energetics-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

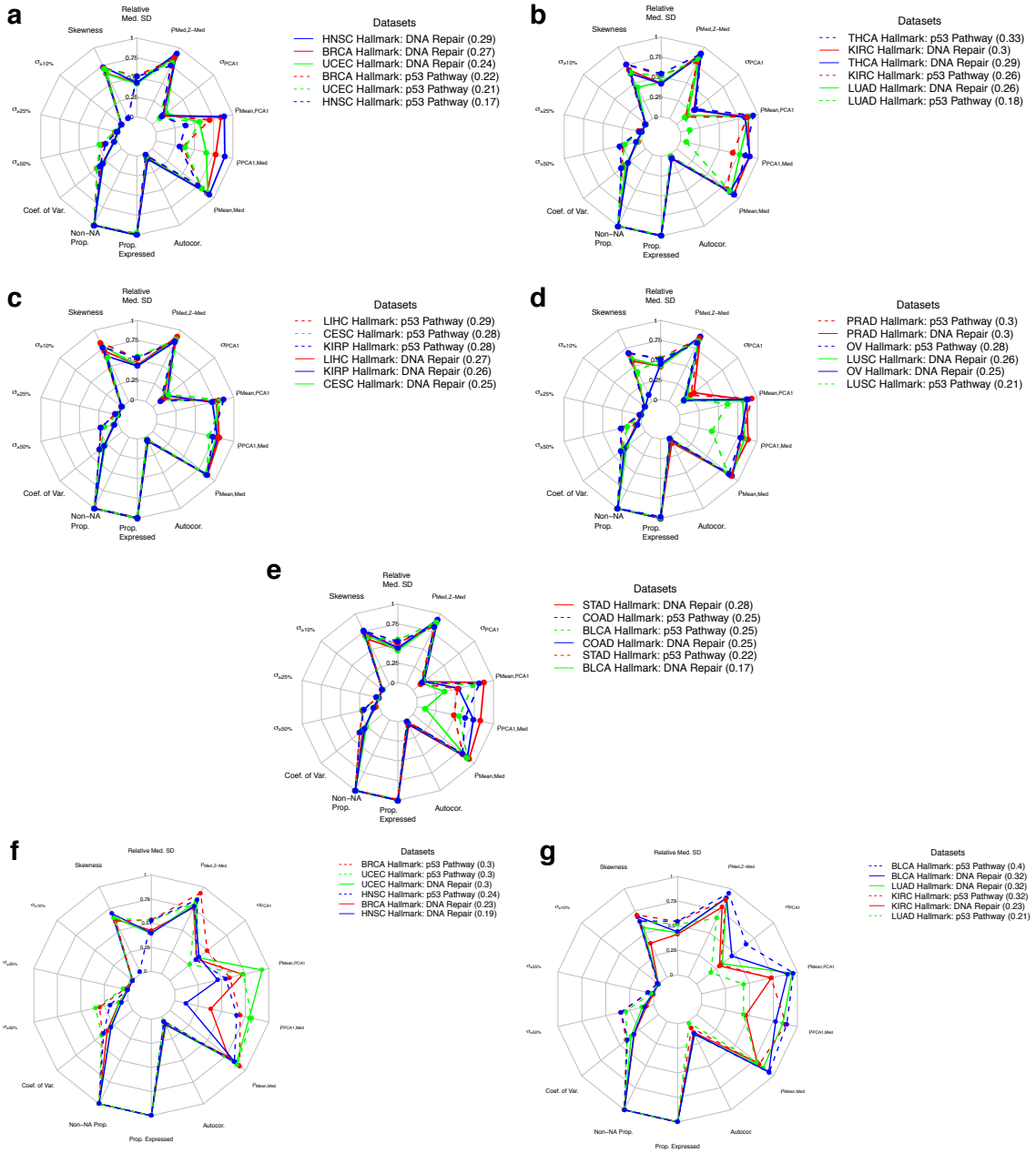

Figure B.4: ***sigQC* radar plots for genome instability-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

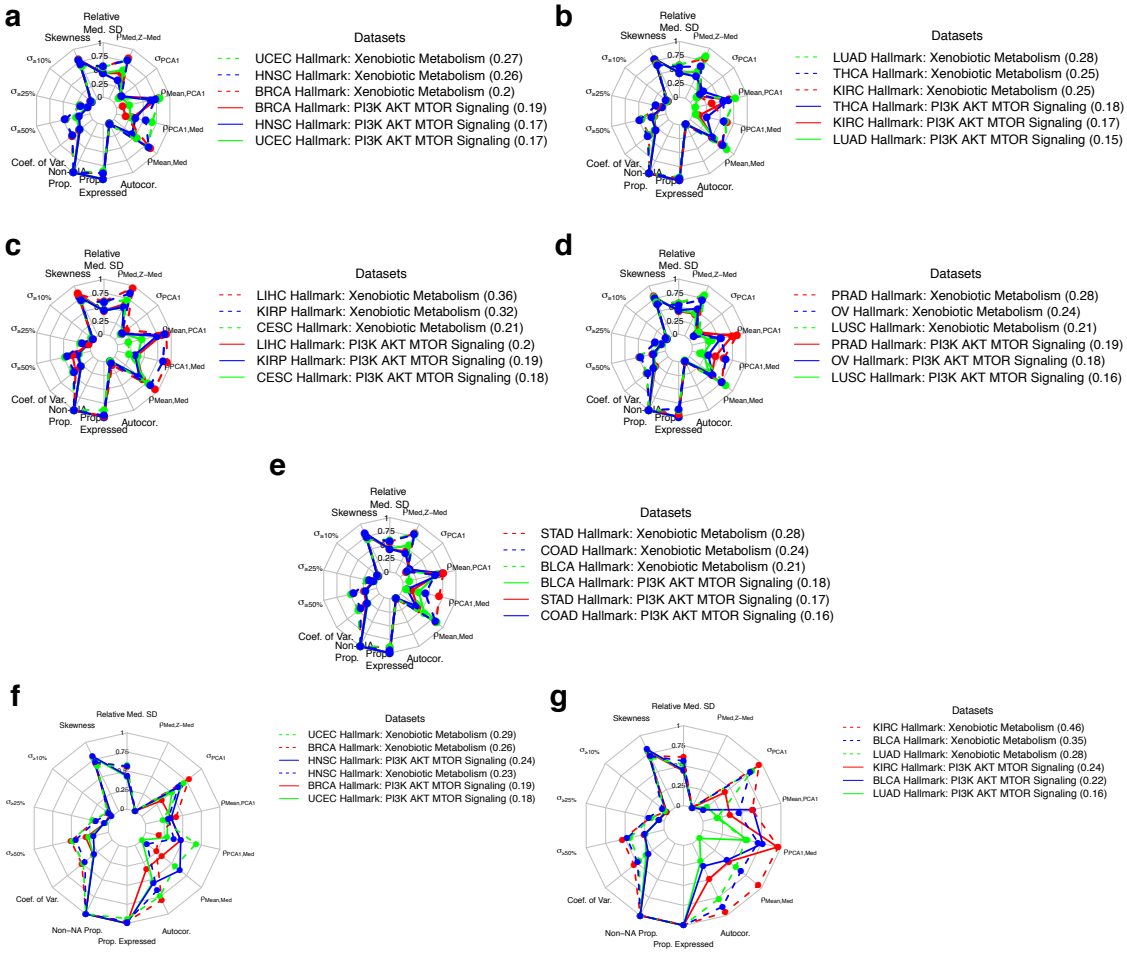

Figure B.5: ***sigQC* radar plots for growth suppressor-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

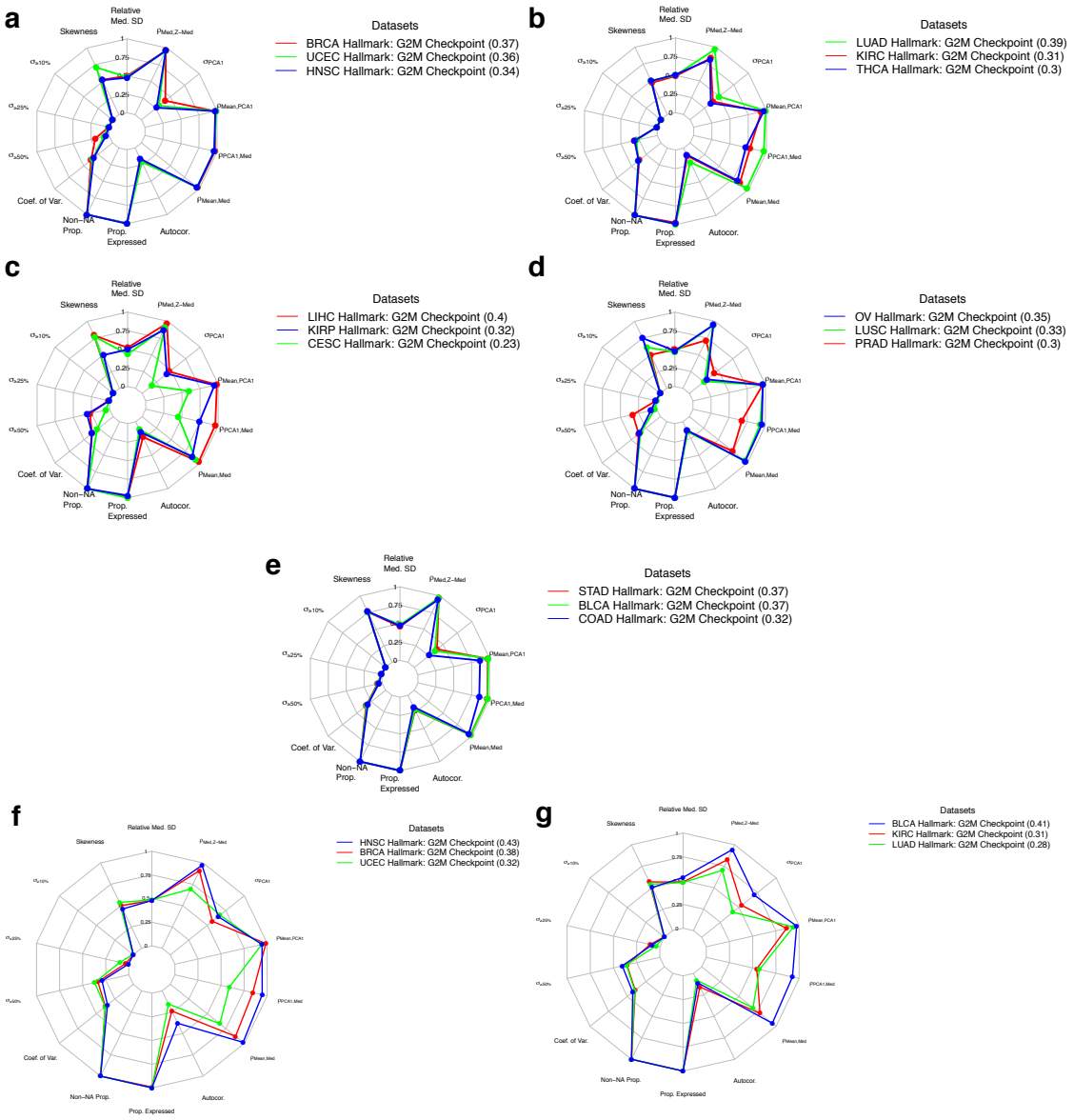

Figure B.6: ***sigQC* radar plots for immortality-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

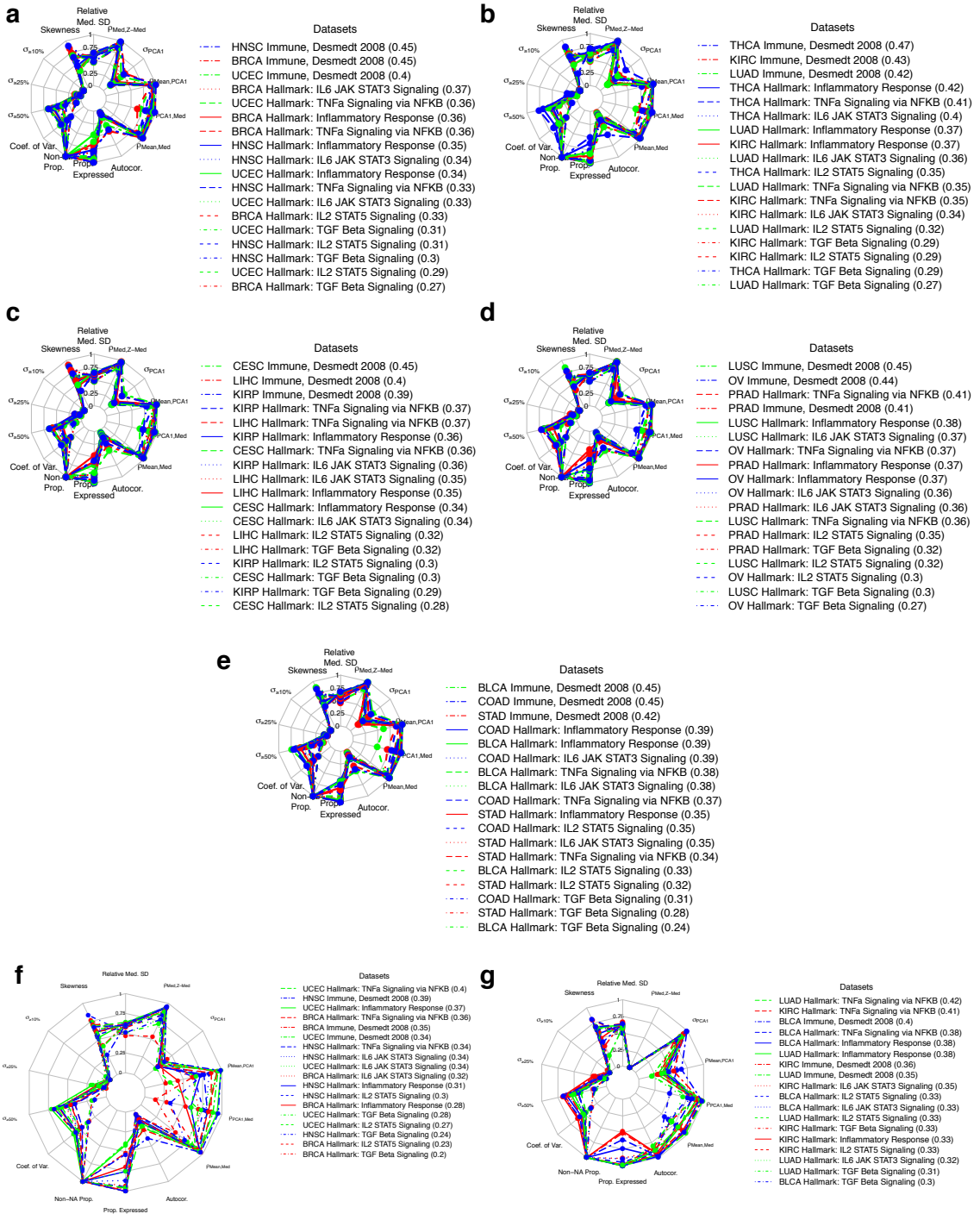

Figure B.7: ***sigQC* radar plots for inflammation-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

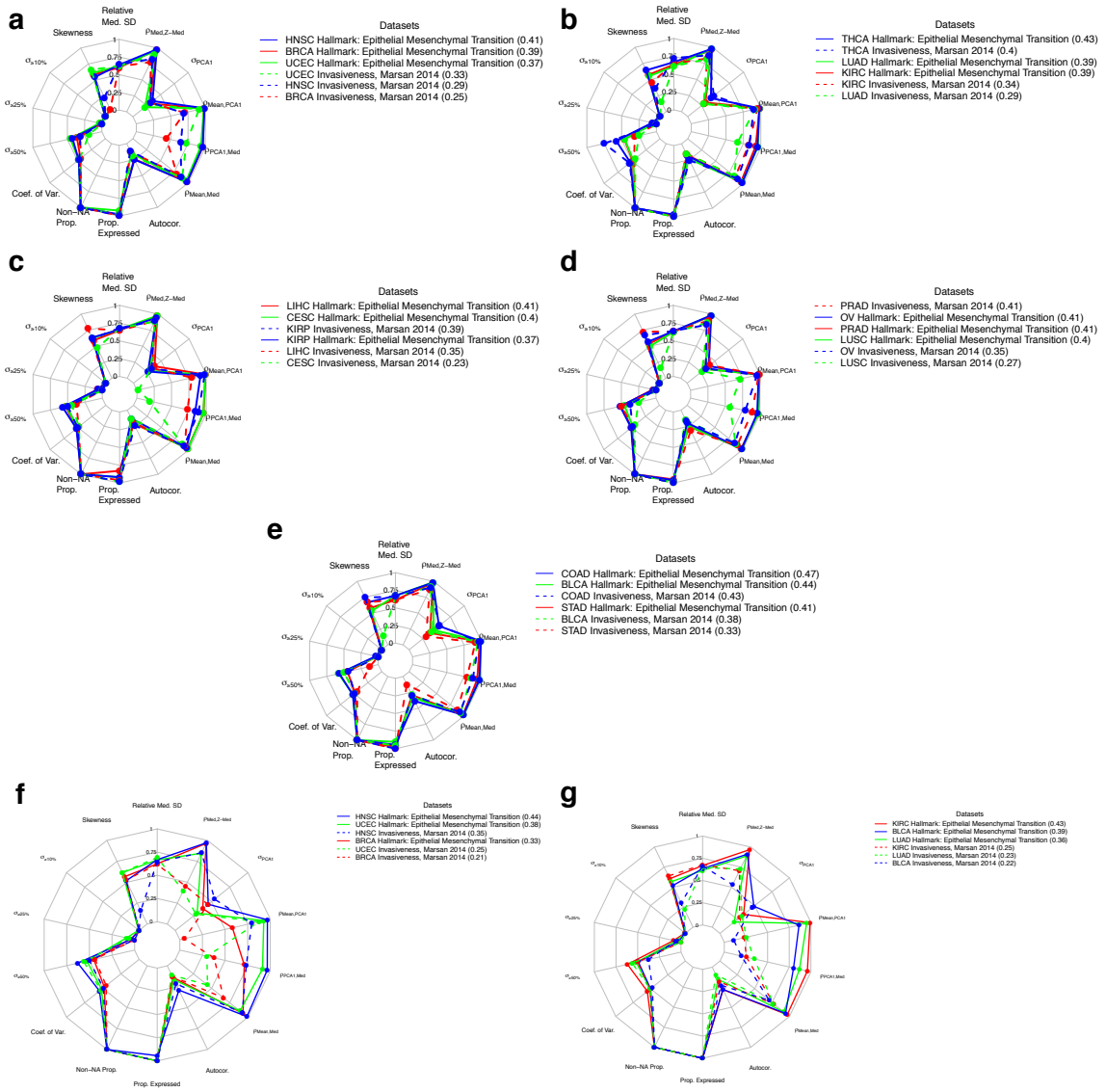

Figure B.8: ***sigQC* radar plots for invasion-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

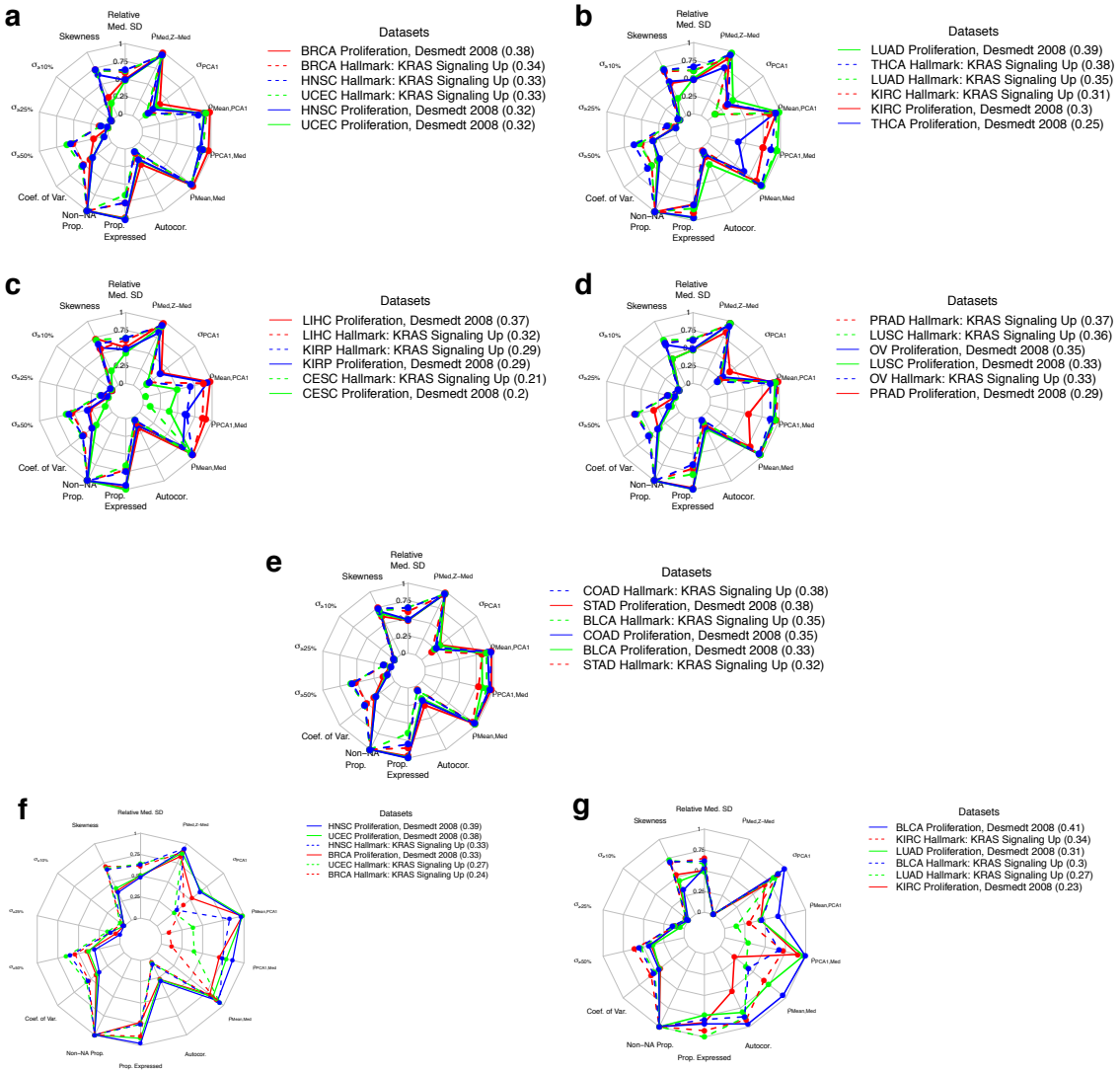

Figure B.9: ***sigQC* radar plots for proliferation-related gene signatures.** (**a**) Shows signatures on BRCA, UCEC, and HNSC datasets, (**b**) Shows signatures on KIRC, LUAD, and THCA datasets, (**c**) Shows signatures on LIHC, CESC, and KIRP datasets, (**d**) Shows signatures on PRAD, LUSC, and OV datasets,and (**e**) Shows signatures on STAD, BLCA, and COAD datasets. (**f**) Shows signatures on BRCA, UCEC, and HNSC adjacent normal datasets. (**g**) Shows signatures on KIRC, LUAD, and BLCA adjacent normal datasets.

B.3 Tables of positively and negatively-associated hallmarks miRNA

Tables providing the miRNA found to be statistically significantly positively associated with each gene signature, and their corresponding rank product statistic p-values (Bonferroni corrected p-value < 0.05) and proportion of false positive values (false positive rate < 0.05) can be found in the supplementary .zip file in the signature_associated_miRNA / miRNA_up subfolder. Likewise, the downregulated miRNA and associated tables can be found in the signature_associated_miRNA / miRNA_down subfolder. The supplementary .zip file may be downloaded from `https://github.com/andrewdhawan/miRNA_hallmarks_of_cancer/`.

B.4 TSG mutation status and associated miRNA expression

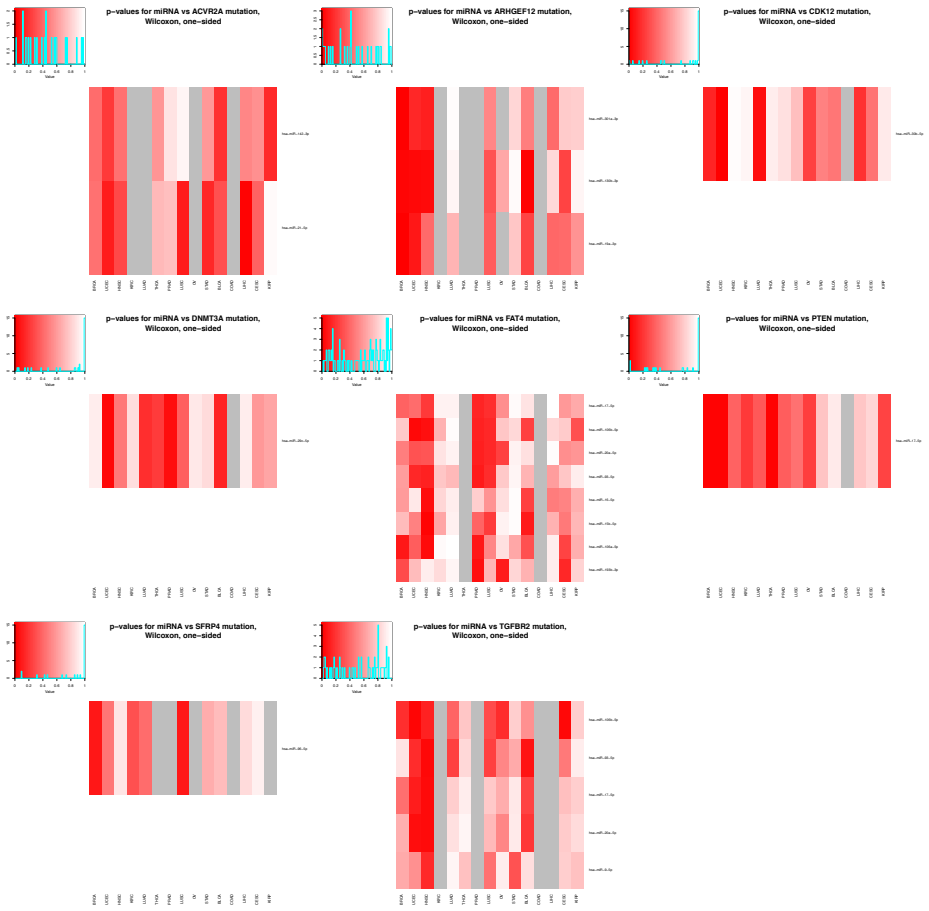

Figure B.10: **miRNA levels tend to decrease variably across cancer types, depending on TSG mutation status.** Heatmap of the p values obtained for the Wilcoxon rank-sum test, one-sided, comparing TSG-associated miRNA expression for statistically significant interactions across cancer types, for TSG mutated and wild-type cases, across tumour types (alternative hypothesis miRNA expression greater in wild-type cases). NA values are indicated in grey.

B.5 Rank product tables, autocorrelation heatmaps for negatively correlated miRNA, methylation probes, and mutations in TSG

Rank product tables, autocorrelation heatmaps for predictors of positive and negative expression for each of the 8 TSG considered may be found in the subfolder TSG, then organised by gene, in the supplementary .zip file, available for download at `https://github.com/andrewdhawan/miRNA_hallmarks_of_cancer/`. All of the probesets and mutation types considered for each of the tumour suppressor genes, are listed in Tables B.1 and B.2, below.

In addition, the ensuing pages contain the figures summarising the analysis for the evidence of exclusive statistical association of TSG by miRNA, when compared with miRNA, methylation, copy number, and mutation. Briefly, as described in the methods, and in Figure 3.6 of the main text, a linear modelling scheme was used, which encompassed each of the variables for methylation levels, miRNA levels, mutation occurrence, and copy number as predictors of TSG mRNA expression across cancer types.

Here, an example is described of how each of the results in the further supplementary figures was interpreted. Results are contrasted for two opposing cases; that of *ACVR2A* in stomach adenocarcinoma, which does not show evidence of exclusivity in association, and *PTEN* in bladder cancer, which does show strong evidence for exclusivity. First, recall that these two TSG are among a set of 8 TSG which appeared to have evidence of miRNA-mediated regulation across cancer types. Having obtained this set of genes, the next step was to determine whether the regulation by miRNA was occurring in the absence or presence of other modes of regulation, such as methylation, copy number aberration, or mutation. As such, a linear modelling paradigm was defined, as described above and in the Methods. Aggregating the coefficients obtained from the linear model across cancer types, it was possible to gain an understanding of how each mode of regulation affected TSG mRNA levels, and which were important relative to the others, as positive and negative regulators. Thus, the negative coefficients in the linear model defined a set of miRNA and methylation probes, as well as the impact of mutations that would function to reduce the level of TSG mRNA.

With these lists of miRNA, methylation probes, and mutation types that were shown to negatively associate with TSG mRNA levels across cancer types, next the

co-occurrence of each of these with each other was determined, which led to the analysis for co-correlation between methylation probe expression levels, miRNA expression, and the variables for mutation occurrence. As contrasted in the sample case considered in Figure B.11a below, distinct patterns emerged in some cases, such as *PTEN* in BLCA, suggesting that methylation probes were co-expressed with each other, and rarely with miRNA, and vice versa, thus leading to the hypothesis of exclusivity of association. In contrast, the case of *ACVR2A* in STAD showed no such separation in Figure B.11a, with many statistically significantly positively correlated miRNA and methylation probes, and no clear exclusivity.

To test this hypothesis rigorously, a resampling-based approach was devised to examine how statistically significant the exclusivity of the regulation of each tumour suppressor gene was. By resampling sets of miRNA and methylation probes, and examining how exclusive the regulation was in each of these cases, the significance of how exclusively occurring these associated entities were, in the cases of interest, could be determined, as depicted in Figure 3.7 in Chapter 3, and summarised in Figure B.11b below for the example case. This shows how, as expected, the exclusivity of regulation observed for *ACVR2A* in STAD is not statistically significant, but it is highly statistically significant in the case of *PTEN* in BLCA.

After the significance of the exclusivity of association of each TSG by each potential mode of regulation, was determined, subgroups of samples for each negatively associated putative regulator were defined, either taking the median level of the associated miRNA, associated methylation probes, or the presence of a mutation to define subgroups. The relationship of each of these subgroups to the level of TSG expression was examined, and this revealed that in some cases, the effect of each of these modes of regulation was indeed statistically significant and important in reducing the level of TSG expression. This is summarised for the example cases shown in Figure B.11c, below, where it is observed how the expression of *ACVR2A* in STAD and *PTEN* in BLCA changes in association with the subgroups of patients defined by the presence or absence of increased methylation, miRNA, or mutation. These results highlight the phenotypic differences between these genomically-distinct subgroups. Further analysis described in Appendix B10 describes how a differential expression analysis was done to further compare transcriptomic similarities and differences between these subgroups as well.

Gene	Methylation probes
PTEN	cg01228636, cg01354923, cg02261018, cg02307823, cg03214660, cg03236184, cg03588460, cg03891929, cg04059318, cg04582473, cg04616691, cg04638773, cg04707787, cg04738091, cg04824711, cg05947570, cg06466203, cg06731059, cg06947206, cg07263825, cg07655693, cg08363193, cg08602305, cg08859916, cg08960754, cg08995089, cg09472211, cg09528884, cg09550257, cg10041390, cg10205334, cg10930218, cg12005026, cg13528847, cg13885325, cg15412736, cg16404460, cg16443434, cg16686761, cg16687447, cg17083429, cg17114151, cg17489897, cg17557106, cg18141918, cg18384060, cg18665732, cg18819818, cg18953873, cg19358349, cg19634213, cg19659388, cg20849549, cg21573601, cg22564317, cg23149470, cg23753021, cg25452974, cg26090855, cg26127345, cg27084903, cg27299538, cg27422496
ACVR2A	cg00532455, cg02093647, cg03601011, cg06120425, cg06907069, cg07526221, cg07969095, cg12968518, cg14420245, cg14425722, cg14689355, cg14926149, cg16081228, cg17174566, cg21233506, cg21579828, cg22464182, cg23727674, cg27112146
ARHGEF12	cg00395063, cg03274991, cg04919489, cg05099464, cg06767612, cg07099388, cg08318018, cg09754341, cg10407488, cg10493270, cg10738003, cg10952477, cg12819548, cg12823408, cg12851792, cg15028899, cg15143809, cg15690696, cg15892763, cg16106770, cg16757423, cg17030562, cg18567470, cg21468385, cg22276271, cg23618830, cg24566217, cg25242756, cg26098650, cg26625290, cg26681847, ch.11.2495959R
CDK12	cg00061989, cg06862673, cg08170745, cg09102835, cg10398950, cg12424509, cg12477119, cg17557704, cg20708332, cg20936107, cg21037155, cg22133495, cg26279814, cg27544759, ch.17.1008464R, ch.17.1009718F
DNMT3A	cg00050692, cg00220517, cg00277048, cg00856404, cg00886730, cg00898683, cg00912598, cg02118630, cg02208653, cg02746110, cg03314052, cg03463641, cg03766400, cg04058399, cg04436772, cg04683068, cg05516842, cg05544807, cg05652528, cg05896193, cg06112956, cg06224893, cg06748978, cg07150430, cg07720334, cg08316074, cg08485187, cg08493294, cg09986894, cg10142668, cg10239163, cg10270719, cg10525105, cg10614445, cg10616515, cg10749994, cg11343289, cg11354105, cg11779362, cg11798660, cg12066181, cg13076778, cg13344237, cg13558695, cg13828701, cg14189391, cg15150970, cg15302376, cg15843262, cg15990840, cg15998962, cg17137500, cg17207266, cg17742416, cg18889183, cg19256292, cg19489797, cg19862213, cg20303441, cg20669908, cg20702417, cg20948740, cg21598294, cg21629895, cg21708767, cg22705918, cg22731525, cg23009818, cg23042148, cg23393100, cg23569120, cg23903708, cg25044635, cg25096282, cg26470599, cg26544247, cg26803803, cg26995204, cg27369452
FAT4	cg00990763, cg03404279, cg03527919, cg04023369, cg04171487, cg04373334, cg04459504, cg05118638, cg08575049, cg08644023, cg10399929, cg10731073, cg12058185, cg12828819, cg13742182, cg15795630, cg17265829, cg17760043, cg18202623, cg22911422, cg23901852, cg25879360, cg26389756
SFRP4	cg01613122, cg01689311, cg04651042, cg05682561, cg06161814, cg08261094, cg09594069, cg10806140, cg11878069, cg12515638, cg13400306, cg14846368, cg16433922, cg18723937, cg19166347, cg20019546, cg21122375, cg22826141, cg23169784, cg23569180, cg25783719
TGFBR2	cg03256955, cg03420580, cg03630790, cg04916416, cg05074709, cg05450916, cg06270049, cg06784602, cg07285675, cg07613391, cg09417692, cg09668216, cg12419522, cg12541591, cg12926720, cg13504215, cg13724812, cg13859541, cg14910241, cg15171154, cg15270950, cg15724876, cg16299428, cg17546721, cg17786388, cg19408535, cg19482049, cg19615017, cg19995459, cg20216935, cg21814995, cg23485307, cg24321706, cg24719910, cg24952959, cg25438762, cg26376346

Table B.1: **Probe sets for methylation of each tumour suppressor gene considered in analysis. Data used are from TCGA methylation studies across cancer types considered.**

Gene	Mutation types
PTEN	Missense Mutation, Nonsense Mutation, Frame Shift Del, Frame Shift Ins, Splice Site, In Frame Del, In Frame Ins, RNA
ACVR2A	Missense Mutation, Frame Shift Del, Silent, Nonsense Mutation, Frame Shift Ins, In Frame Del, Splice Site
ARHGEF12	Missense Mutation, Nonsense Mutation, Frame Shift Del, Splice Site, Intron
CDK12	Missense Mutation, Nonsense Mutation, Frame Shift Del, Frame Shift Ins, Splice Site, In Frame Del
DNMT3A	Missense Mutation, Frame Shift Del, Splice Site, Nonsense Mutation, Frame Shift Ins
FAT4	Missense Mutation, Nonsense Mutation, Frame Shift Del, Splice Site, Frame Shift Ins, In Frame Del
SFRP4	Missense Mutation, Frame Shift Ins, Frame Shift Del, Splice Site
TGFBR2	Missense Mutation, Nonsense Mutation, In Frame Del, Frame Shift Del, Splice Site, Frame Shift Ins

Table B.2: **Mutation types (non-silent) for each tumour suppressor gene considered in analysis. Data used are from TCGA mutation studies across cancer types considered, as reported by Oncotated calls accessed from the Firebrowse data portal.**

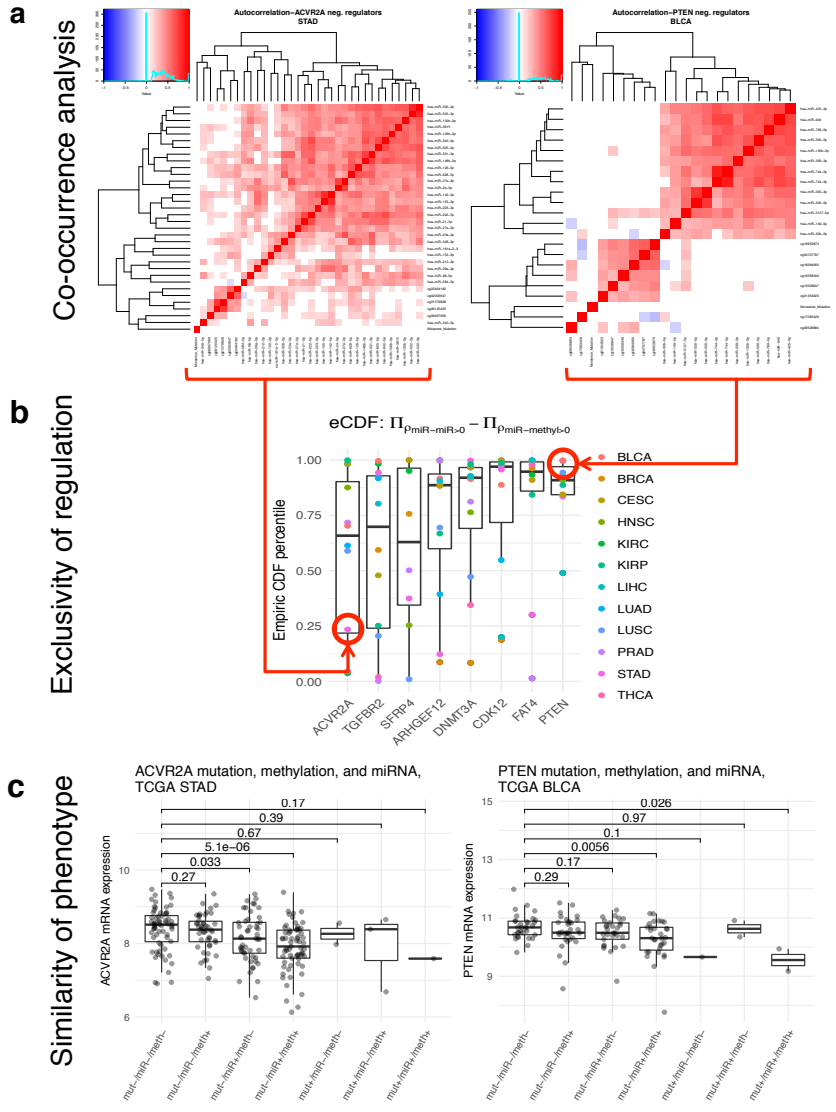

Figure B.11: Sample of analysis done to determine exclusivity of TSG regulation. Example shown is for the two differing cases of *ACVR2A* in STAD (left) and *PTEN* in BLCA (right). (**a**) Shows the heatmaps obtained for the co-correlation between the mutation occurrence, expression of methylation probes, and miRNA negatively regulating each gene, showing the visual differences in the exclusivity of the regulation of these genes. (**b**) Depicts the significance of the exclusivity observed, as computed by a resampling-based approach, highlighted with red circles for *ACVR2A* in STAD and *PTEN* in BLCA. (**c**) Depicts how the expression of the TSG of interest (*ACVR2A* or *PTEN*) varied between subgroups defined by methylation high or low, miRNA expression high or low, or mutation present or not.

B.6 Autocorrelation heatmaps for negative regulators of TSG

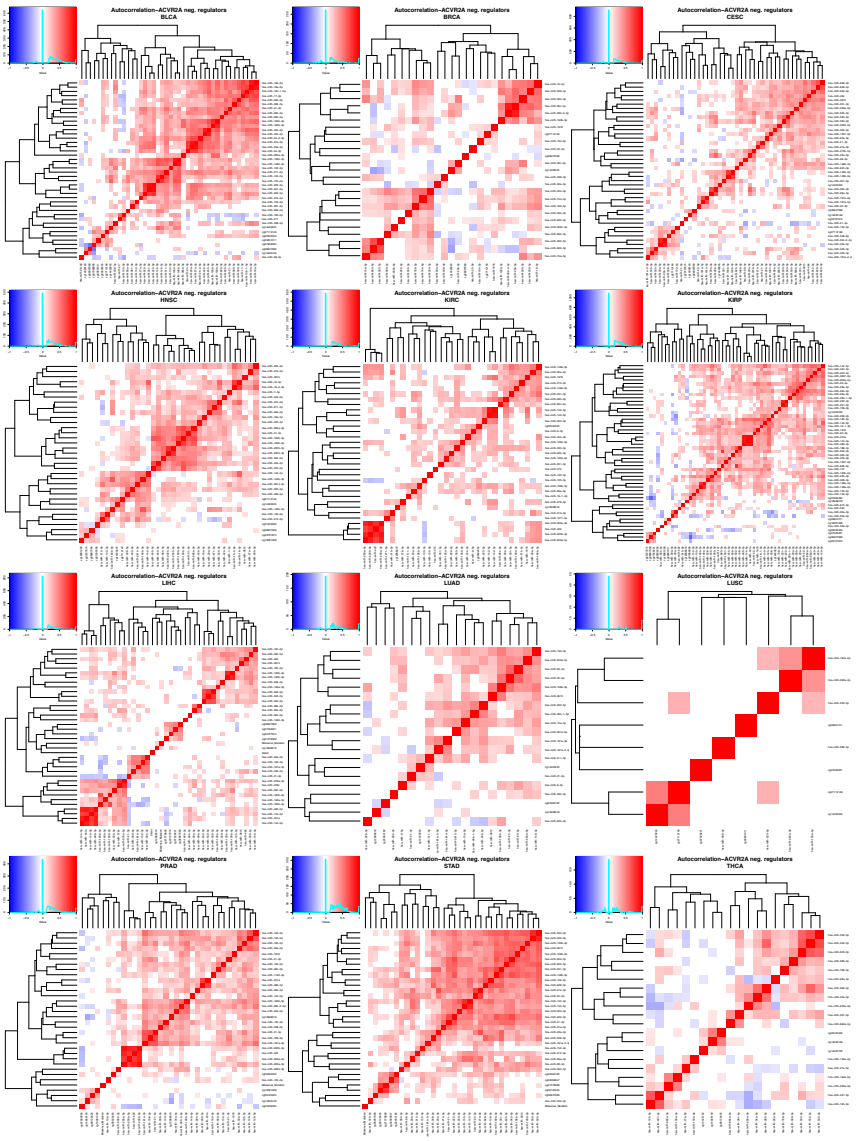

Figure B.12: **Co-occurrence heatmap for negative regulators of *ACVR2A*.** Autocorrelation heatmap for the expression of the identified negative regulators of *ACVR2A*, across the 12 cancer types considered.

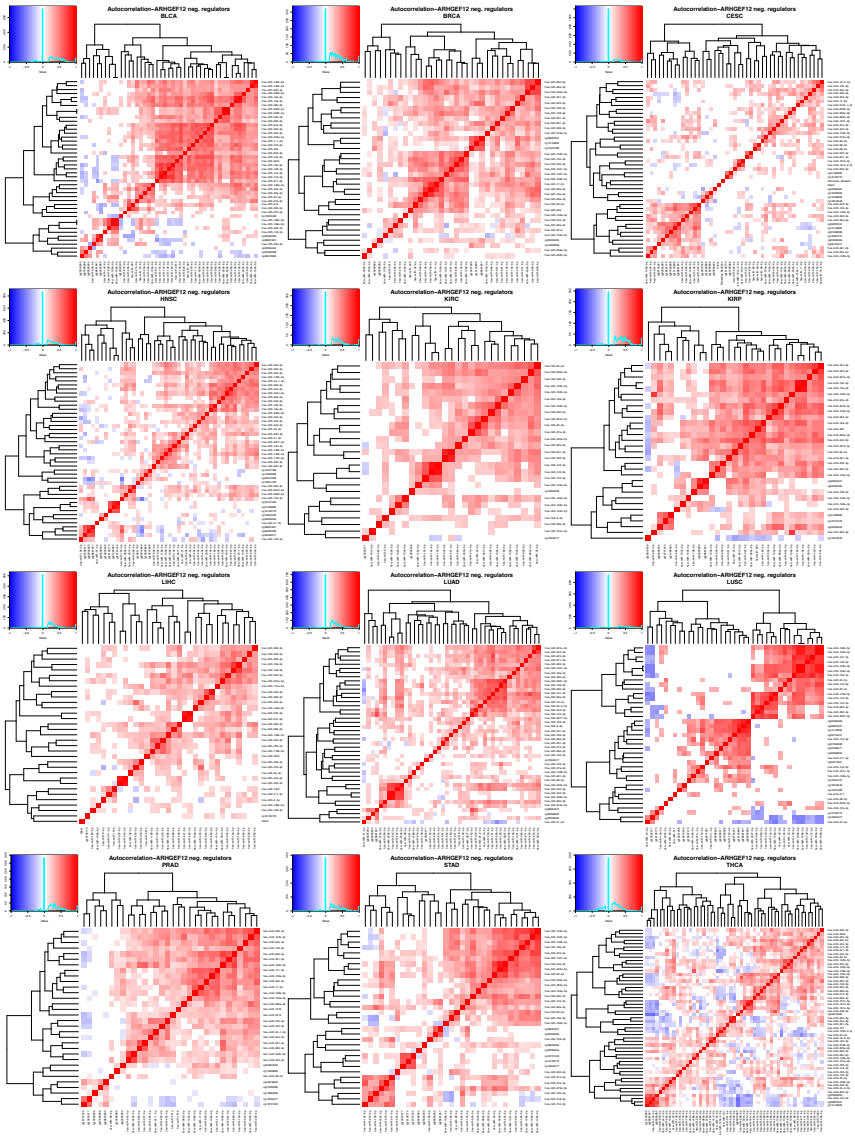

Figure B.13: **Co-occurrence heatmap for negative regulators of *ARHGEF12*.** Autocorrelation heatmap for the expression of the identified negative regulators of *ARHGEF12*, across the 12 cancer types considered.

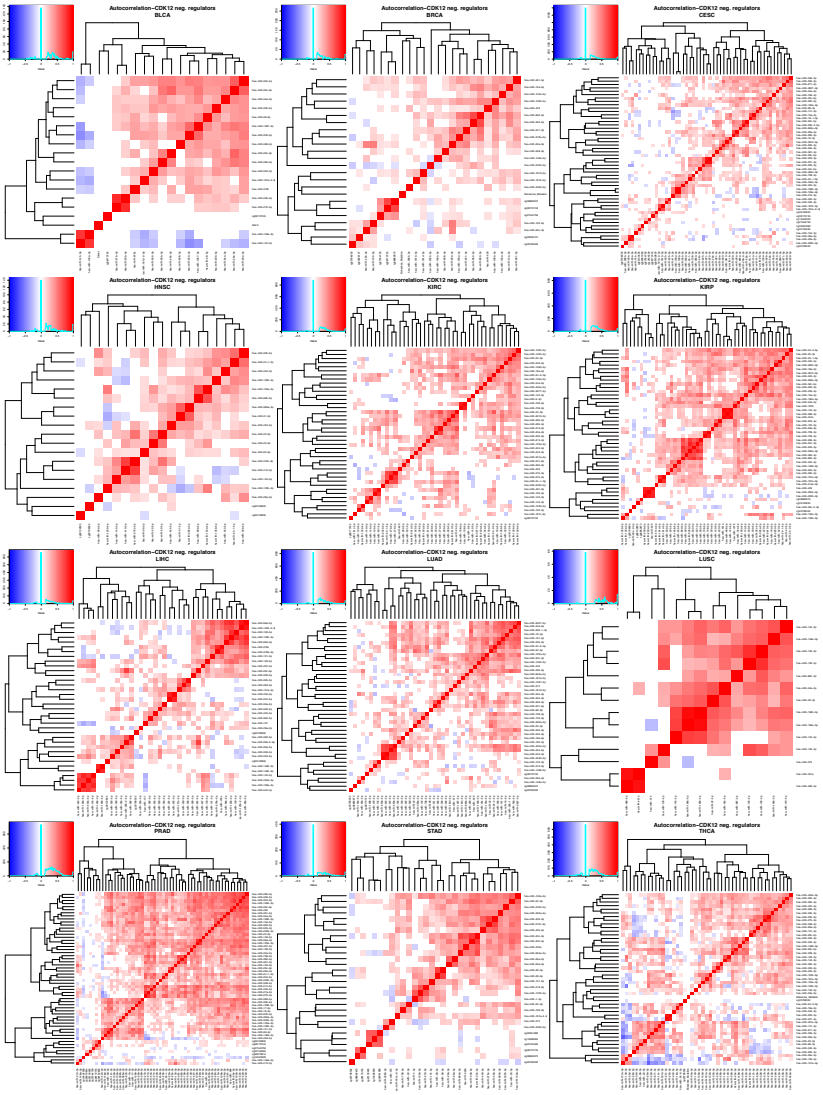

Figure B.14: **Co-occurrence heatmap for negative regulators of *CDK12*.** Autocorrelation heatmap for the expression of the identified negative regulators of *CDK12*, across the 12 cancer types considered.

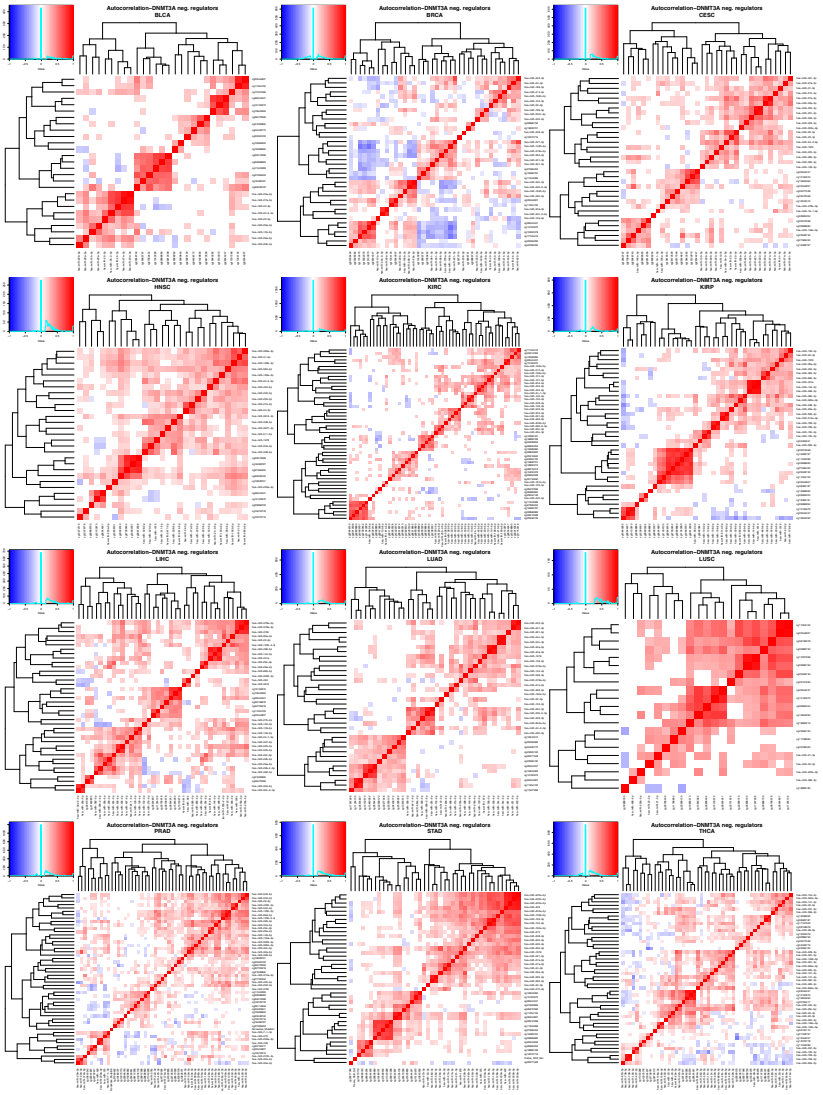

Figure B.15: **Co-occurrence heatmap for negative regulators of *DNMT3A*.** Autocorrelation heatmap for the expression of the identified negative regulators of *DNMT3A*, across the 12 cancer types considered.

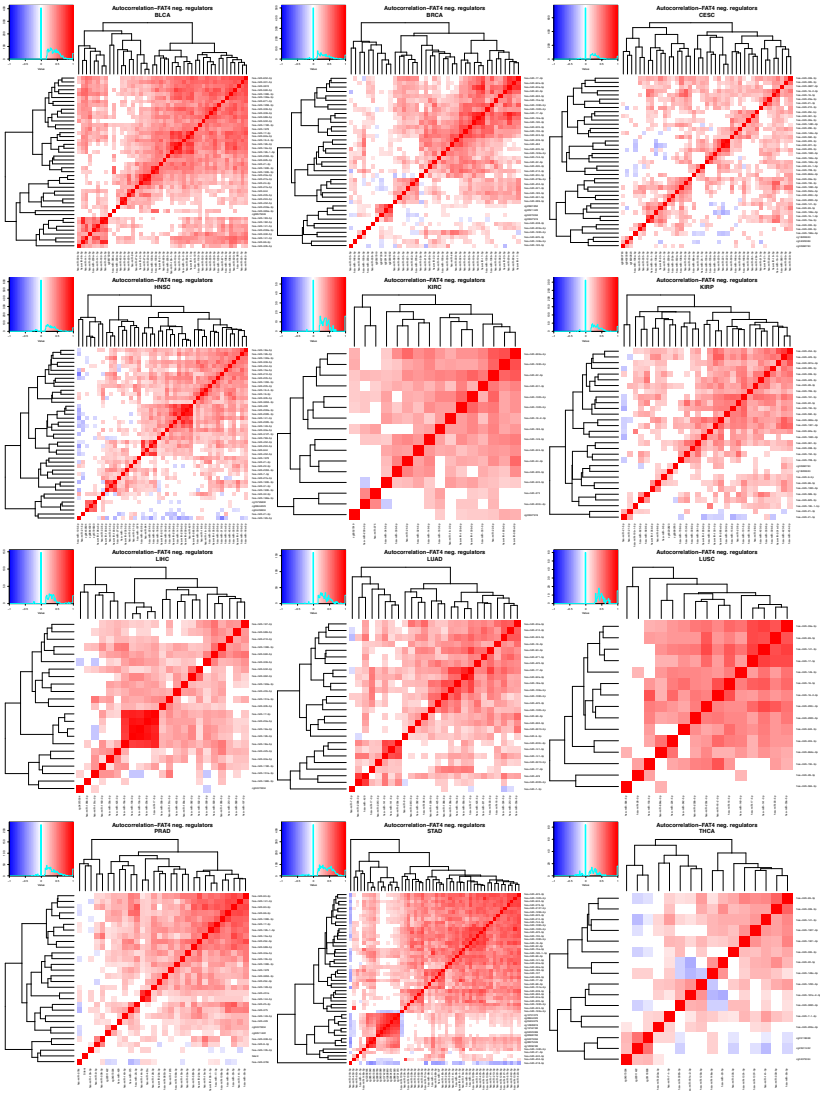

Figure B.16: **Co-occurrence heatmap for negative regulators of *FAT4*.** Autocorrelation heatmap for the expression of the identified negative regulators of *FAT4*, across the 12 cancer types considered.

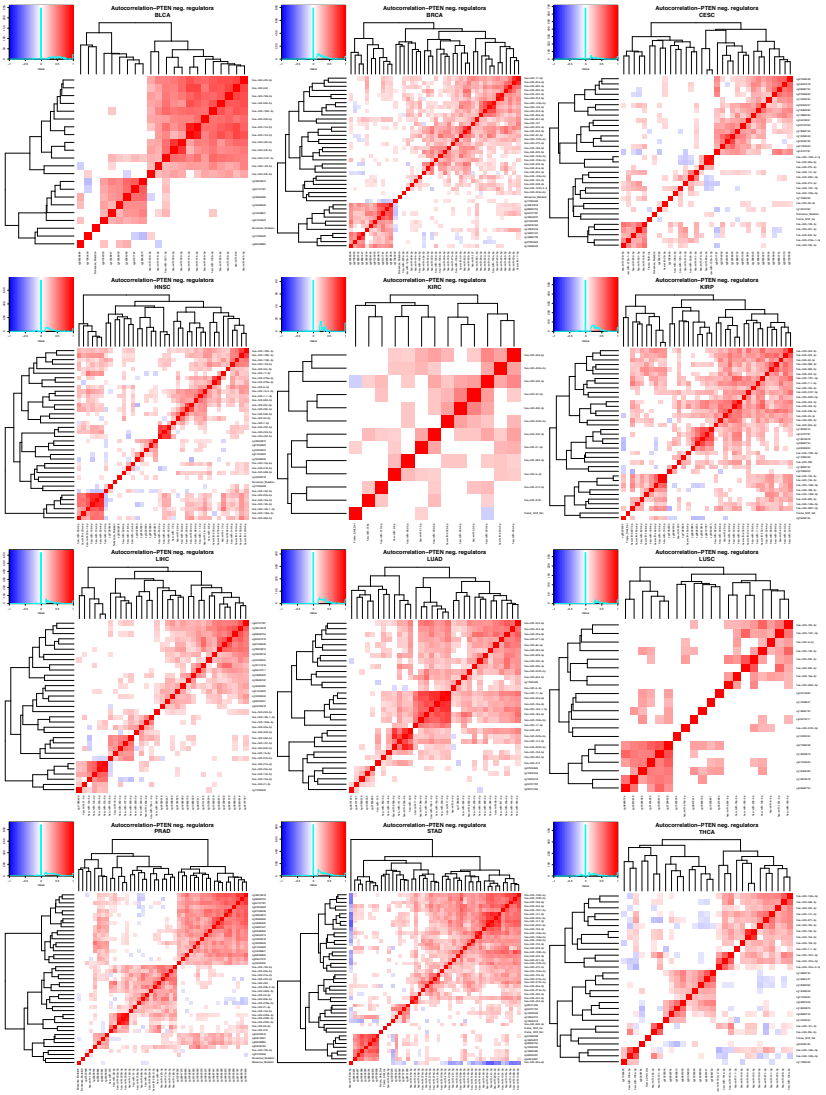

Figure B.17: **Co-occurrence heatmap for negative regulators of *PTEN*.** Autocorrelation heatmap for the expression of the identified negative regulators of *PTEN*, across the 12 cancer types considered.

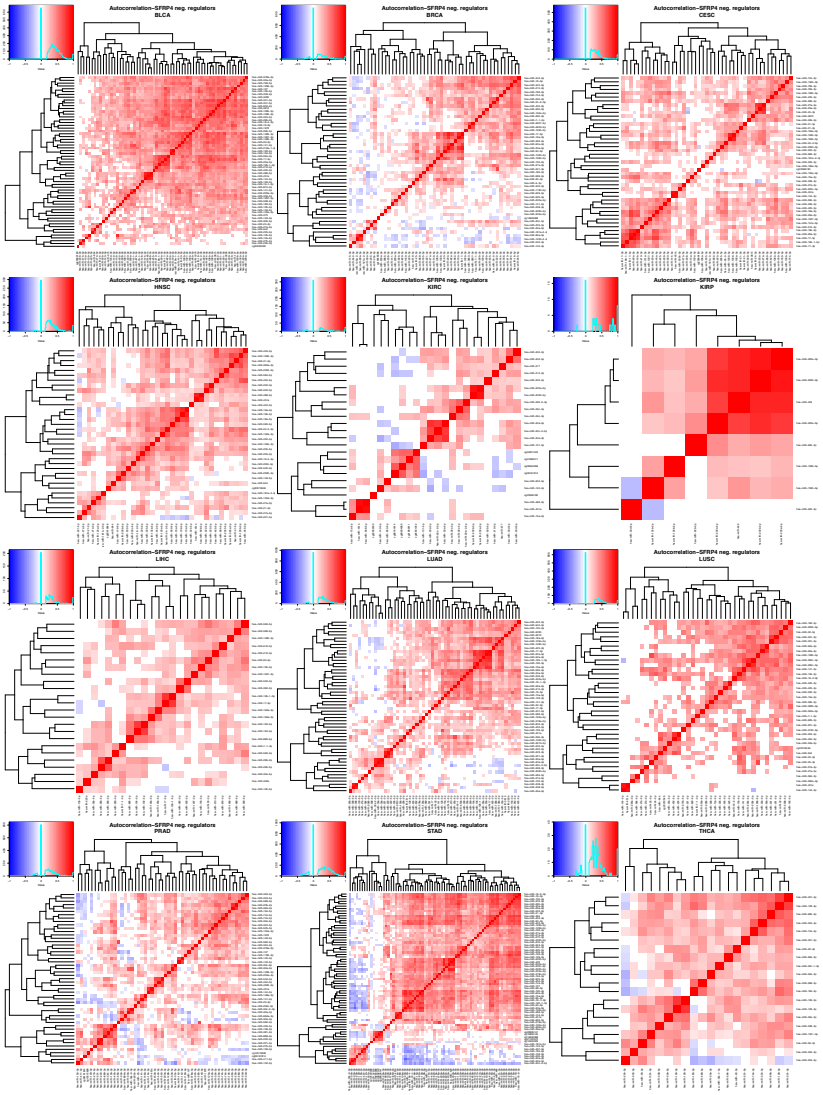

Figure B.18: **Co-occurrence heatmap for negative regulators of *SFRP4*.** Autocorrelation heatmap for the expression of the identified negative regulators of *SFRP4*, across the 12 cancer types considered.

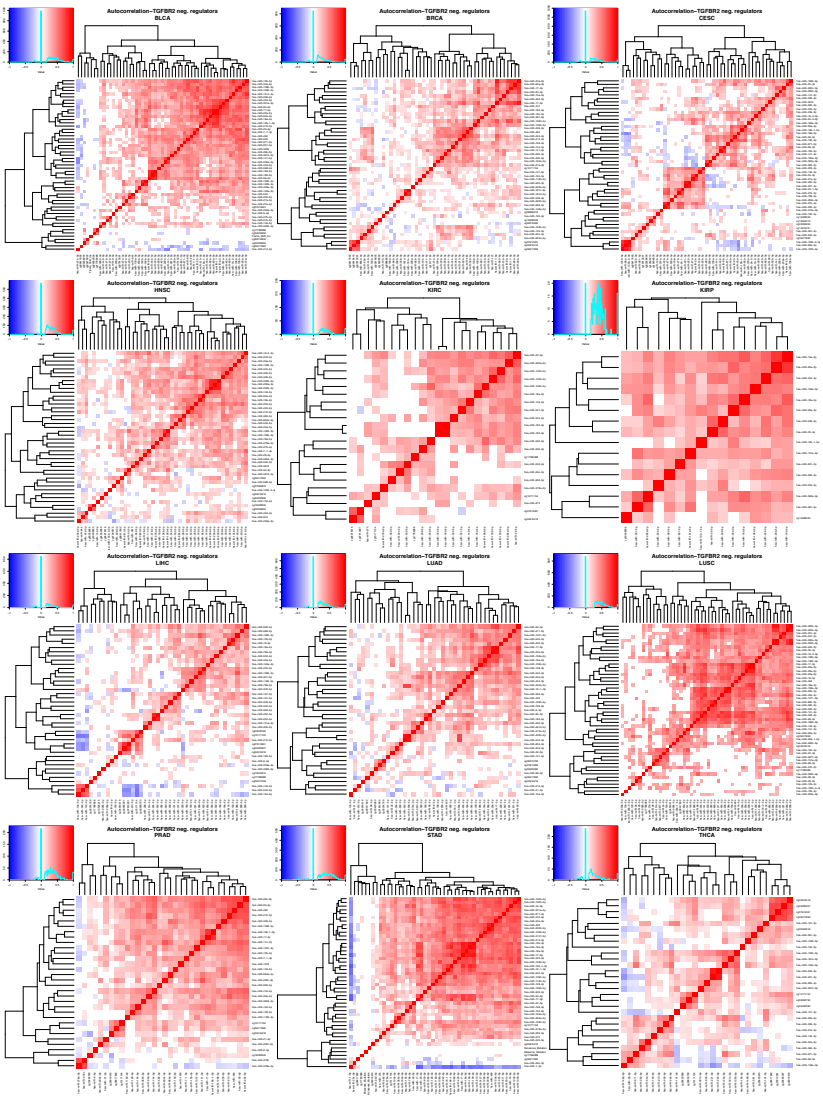

Figure B.19: **Co-occurrence heatmap for negative regulators of *TGFBR2*.** Autocorrelation heatmap for the expression of the identified negative regulators of *TGFBR2*, across the 12 cancer types considered.

B.7 TSG expression and mutation status, miRNA expression, and methylation status

The following pages contain figures representing an analysis first showing the expression of each of the TSG grouped by regulatory type in Figures B.20- B.29. For each tumour suppressor gene considered, the miRNA associated to it were those identified as strong pan-cancer negative predictors of expression, and likewise the methylation probes associated to it were those identified as strong pan-cancer negative predictors of expression. Groups of miRNA high (+) and low (-) were defined by median expression of these TSG-associated miRNA above or below the median for a given set of samples of a particular tumour type. Methylation high (+) and low (-) groups were defined analogously. Mutation groups were defined by the presence (or absence) of non-silent mutations. Cases of TSG and tumour types for which, among the common samples displaying methylation, miRNA expression, mutation, and mRNA expression data, there were fewer than 5 samples with non-silent mutation were excluded from analysis.

Subsequently, in Figures B.30-B.38, the correlation is shown on a scatterplot, for the log2 fold changes for the differentially expressed genes between samples stratified into two groups based on mutation status of the TSG, versus samples with no mutation, lowly methylated sites, and low miRNA expression. Results show that commonly differentially expressed genes among the two groups from the null case of no apparent regulation of the TSG, across cancer types for which at least 5 mutant cases were present, show strong correlation in log2 fold change, suggesting a similar transcriptomic phenotype afforded by methylation and/or miRNA regulation as mutation for these TSG. Differential expression of genes was computed by the EBSeq (Empiric Bayesian) approach to estimate differential expression of genes taken from RSEM non-normalised raw counts for mRNA expression data from the TCGA Firebrowse data portal. In all analyses a false discovery rate of 0.05 was used.

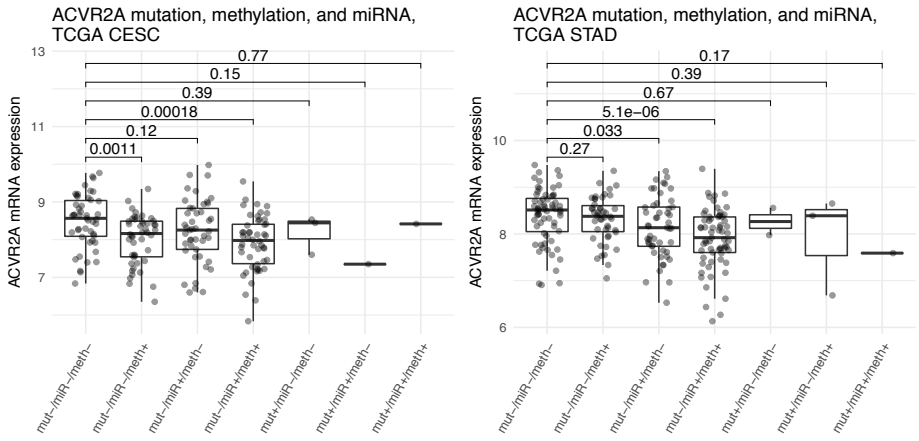

Figure B.20: ***ACVR2A* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *ACVR2A* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

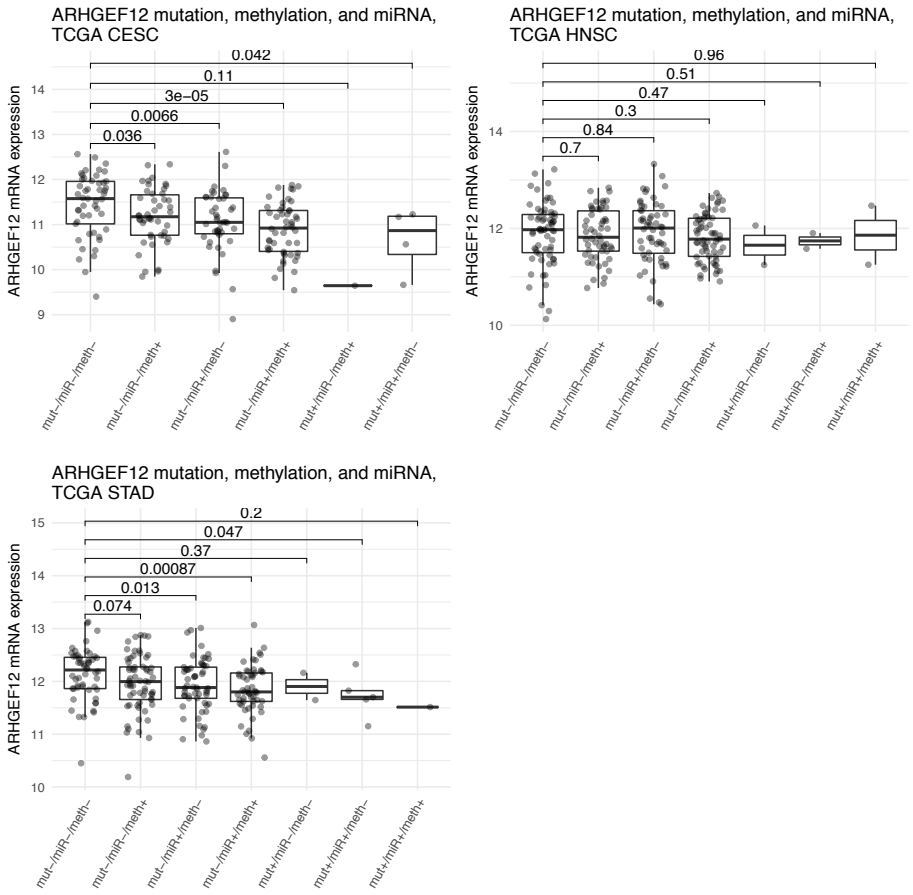

Figure B.21: ***ARHGEF12* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *ARHGEF12* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

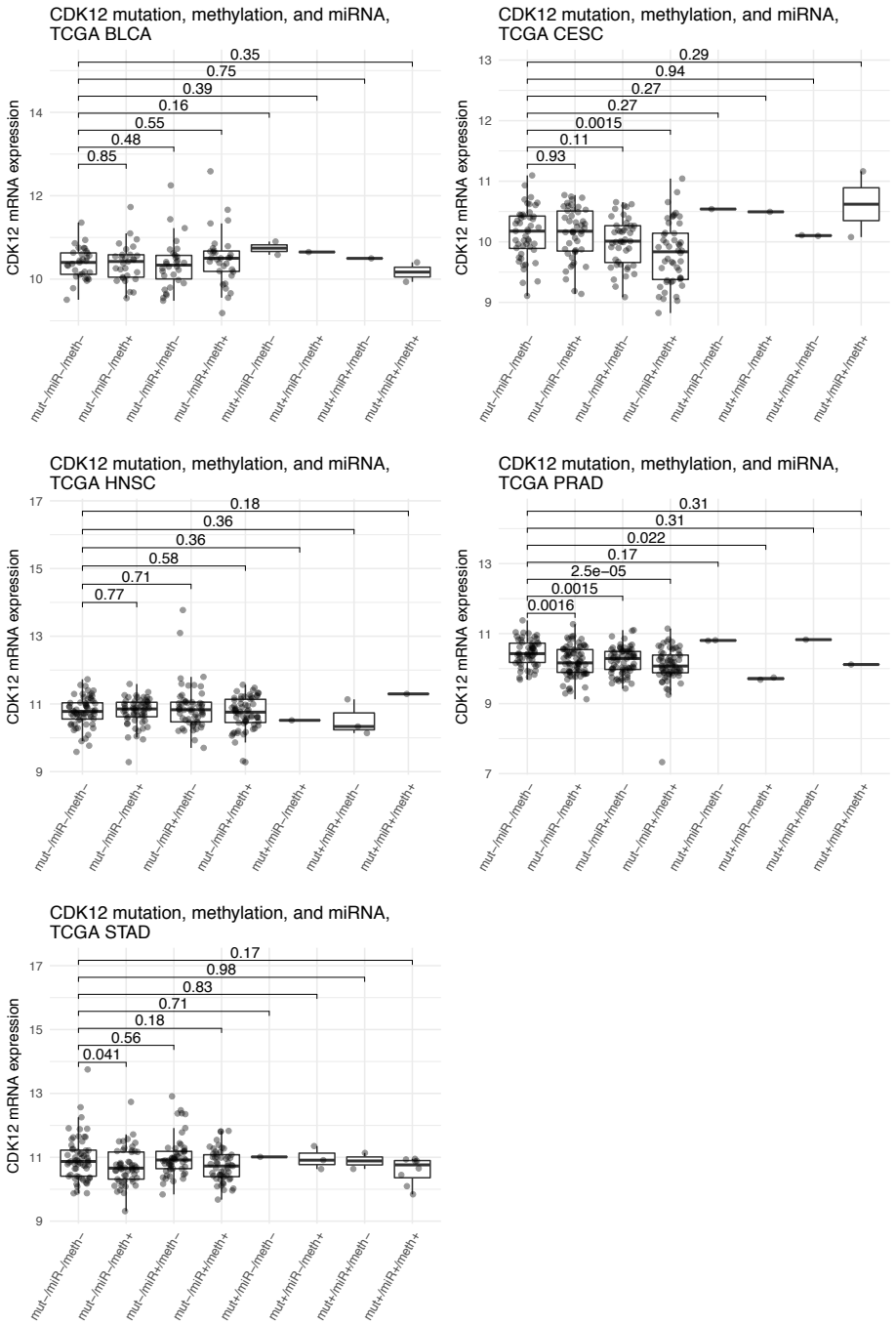

Figure B.22: ***CDK12* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *CDK12* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

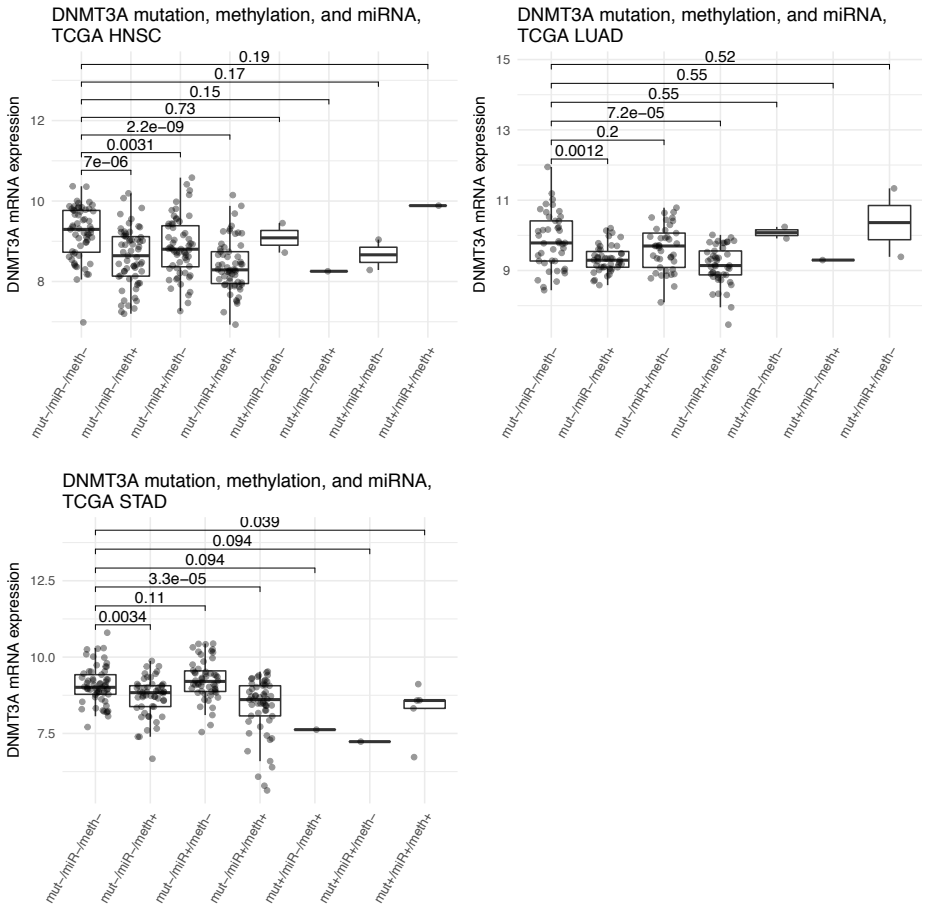

Figure B.23: ***DNMT3A* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *DNMT3A* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

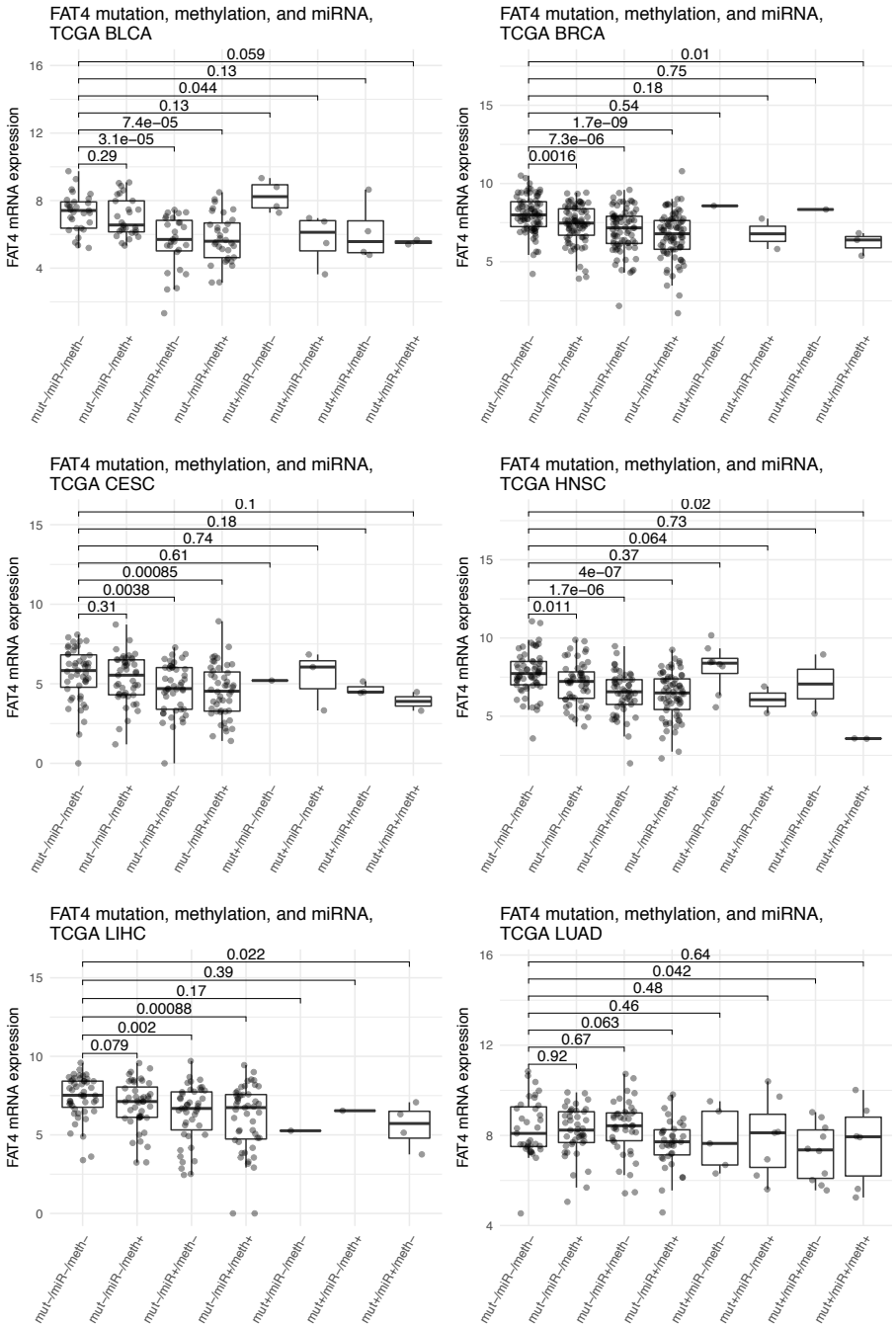

Figure B.24: ***FAT4* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *FAT4* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

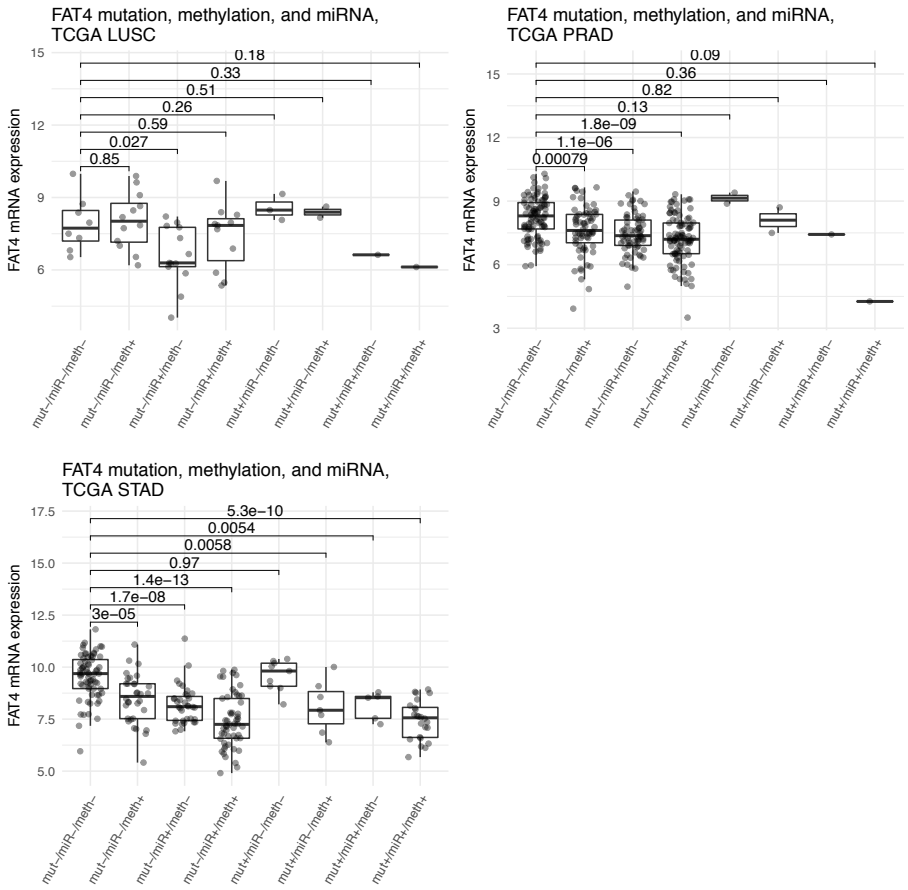

Figure B.25: ***FAT4* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *FAT4* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

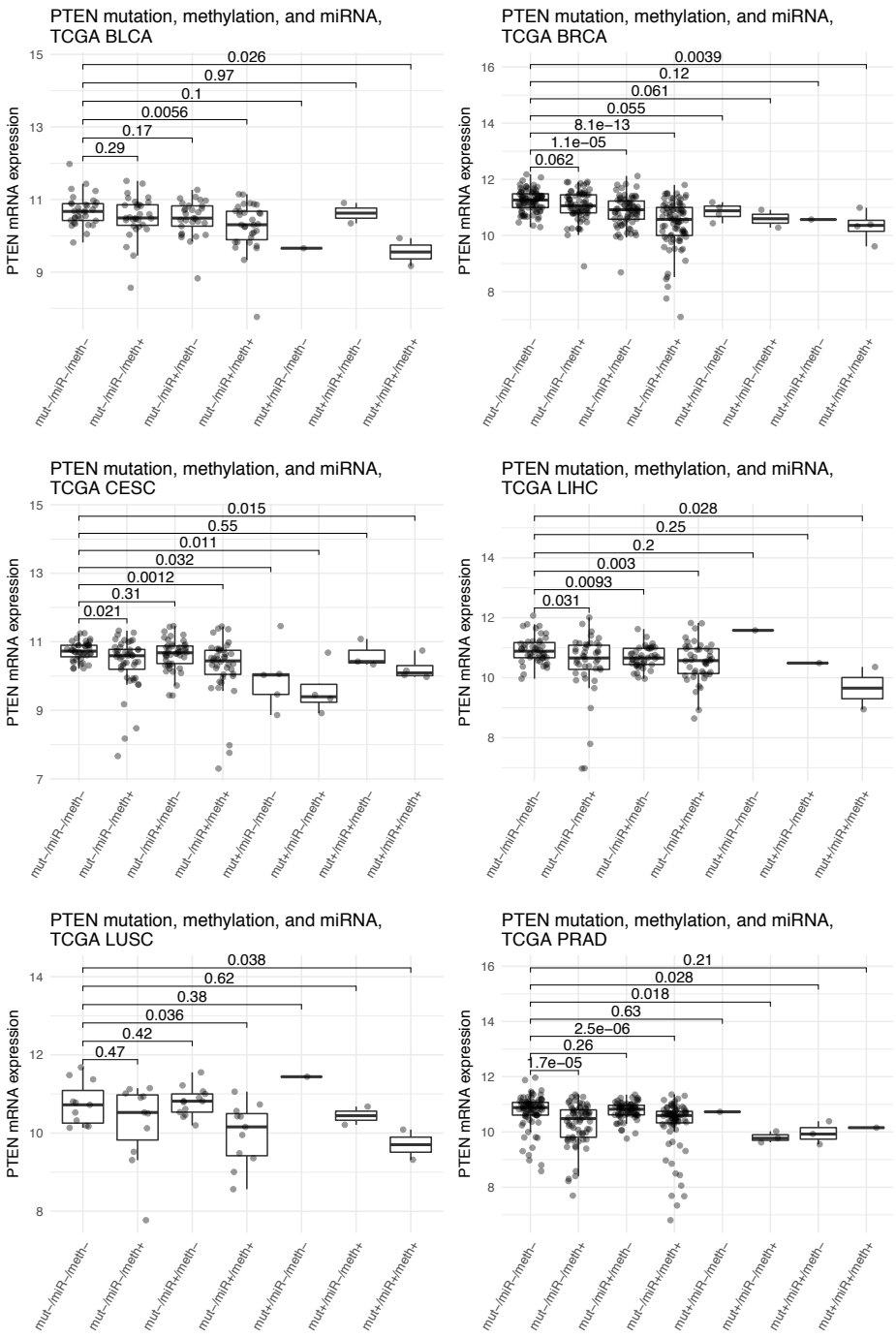

Figure B.26: ***PTEN* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *PTEN* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

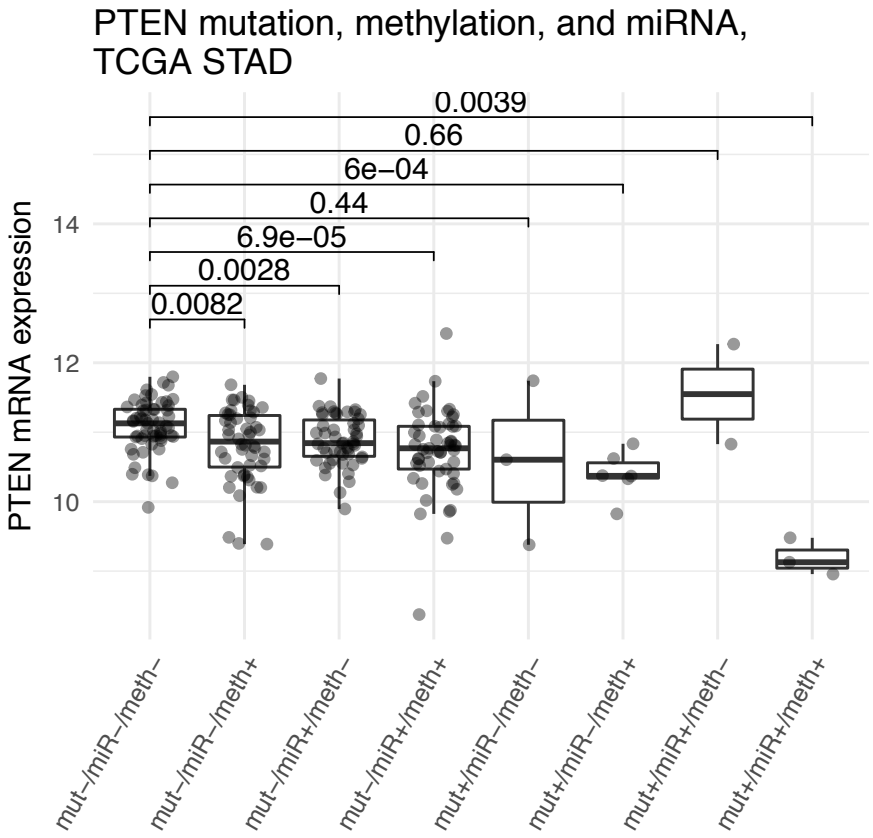

Figure B.27: ***PTEN* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *PTEN* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

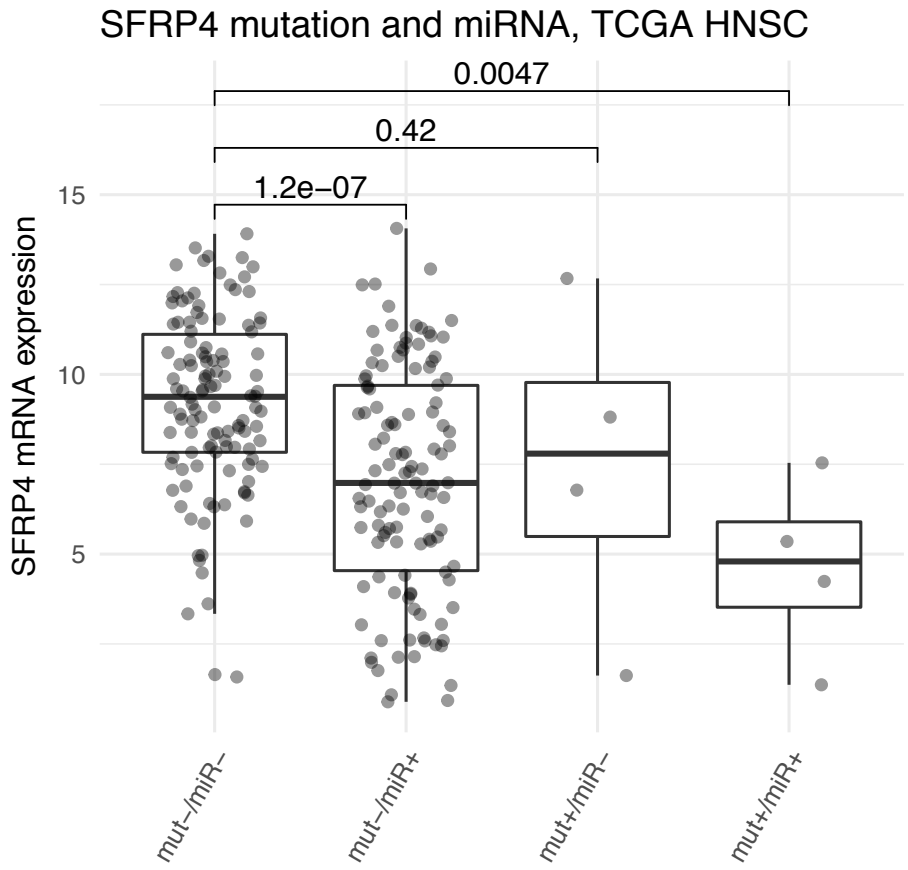

Figure B.28: ***SFRP4* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *SFRP4* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples.

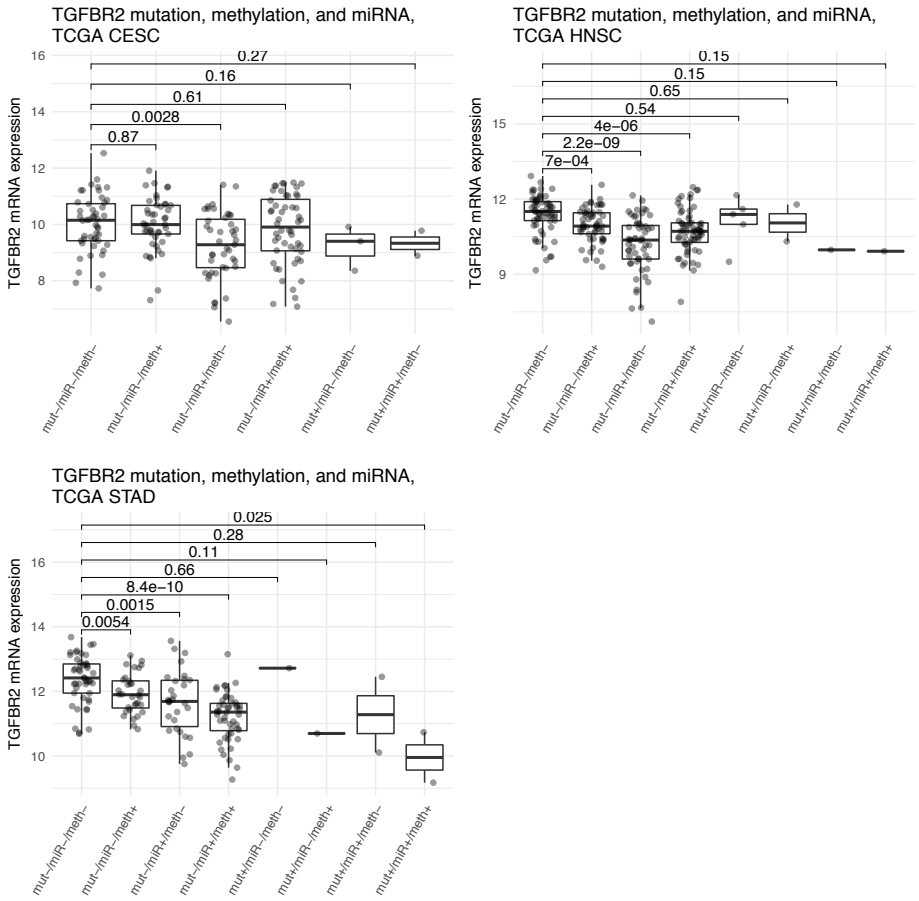

Figure B.29: ***TGFBR2* expression differs between subgroups of negative regulators expressed.** Boxplots showing expression of *TGFBR2* across the distinct regulatory subgroups, across tumour types with at least 5 samples showing non-silent mutation. miRNA status is determined by whether median of identified negatively associated miRNA show expression above or below median value across samples, methylation status is defined analogously.

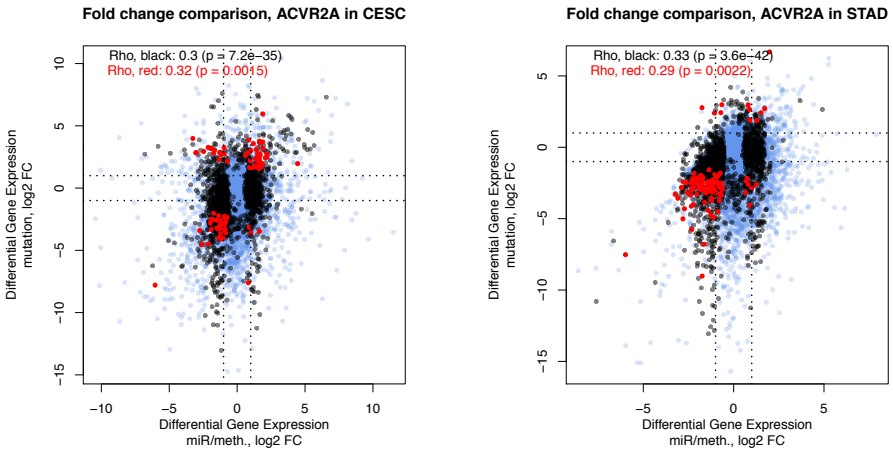

Figure B.30: **A similar profile of differential expression is observed in *ACVR2A* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *ACVR2A* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *ACVR2A* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

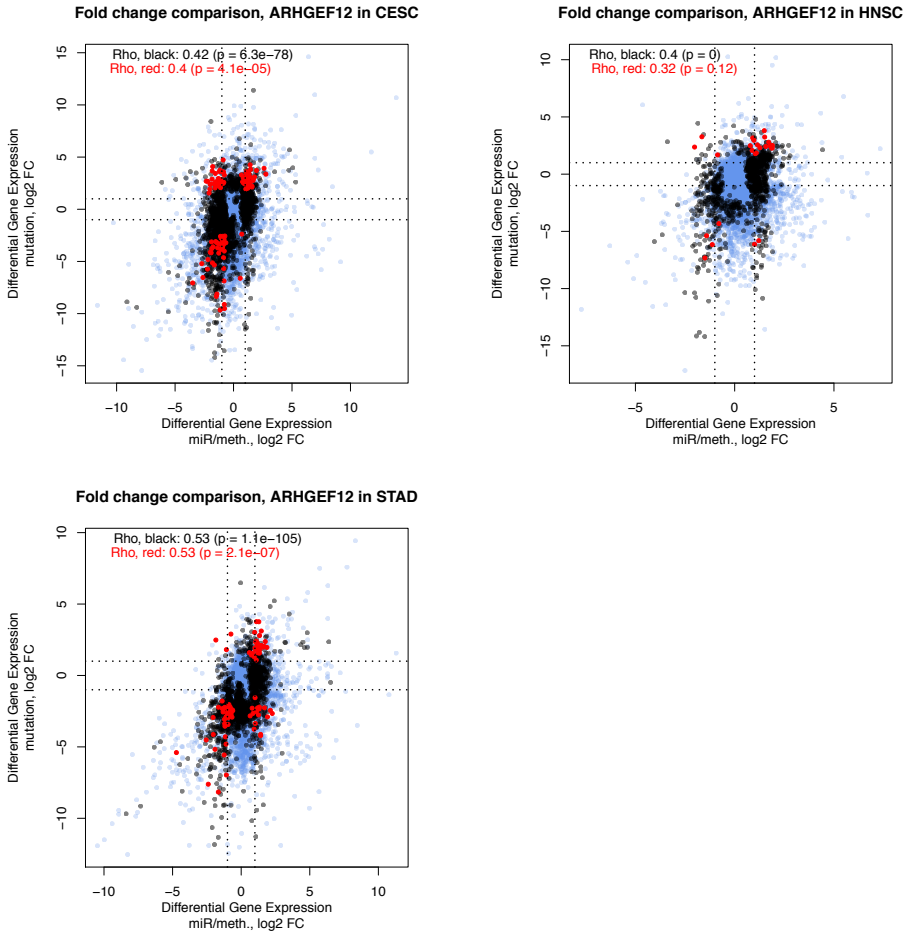

Figure B.31: **A similar profile of differential expression is observed in *ARHGEF12* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *ARHGEF12* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *ARHGEF12* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

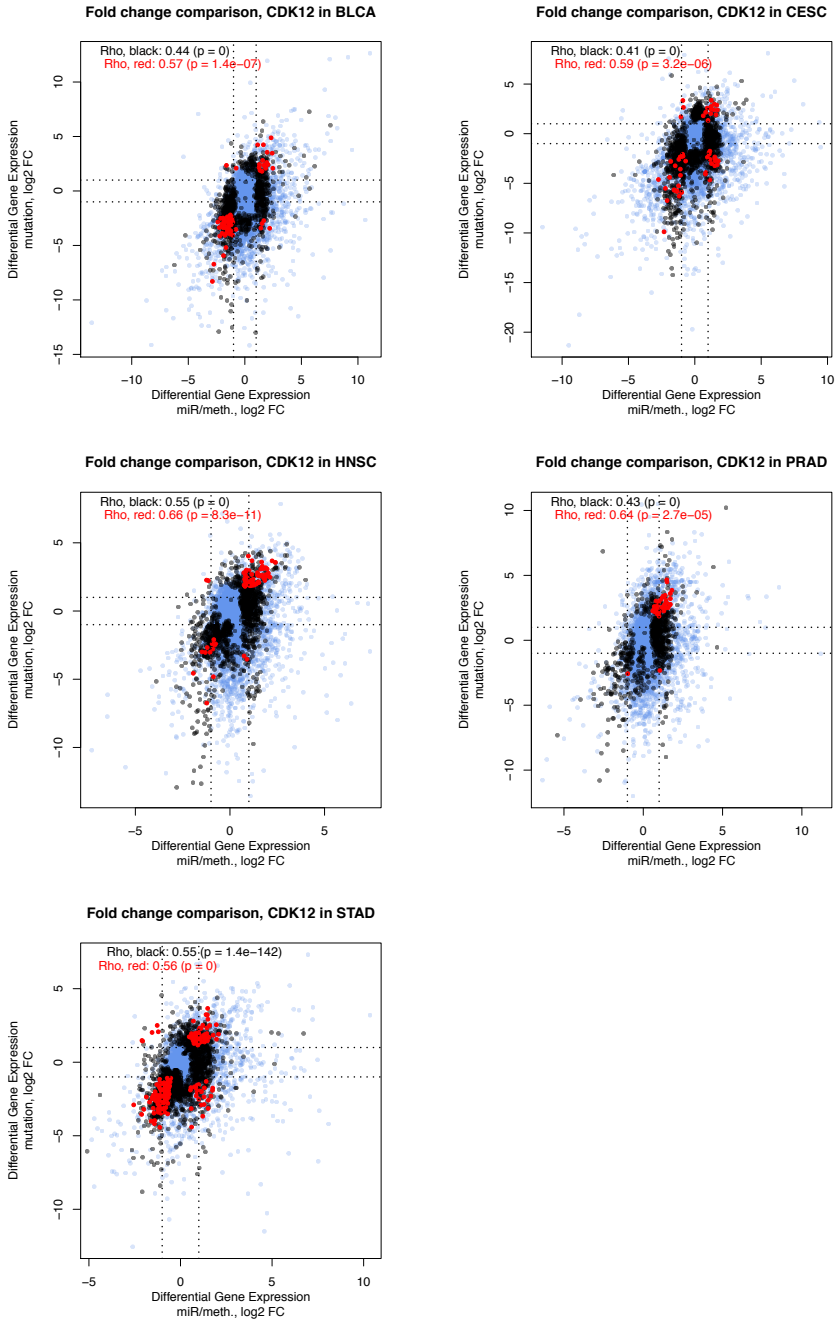

Figure B.32: **A similar profile of differential expression is observed in *CDK12* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *CDK12* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *CDK12* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

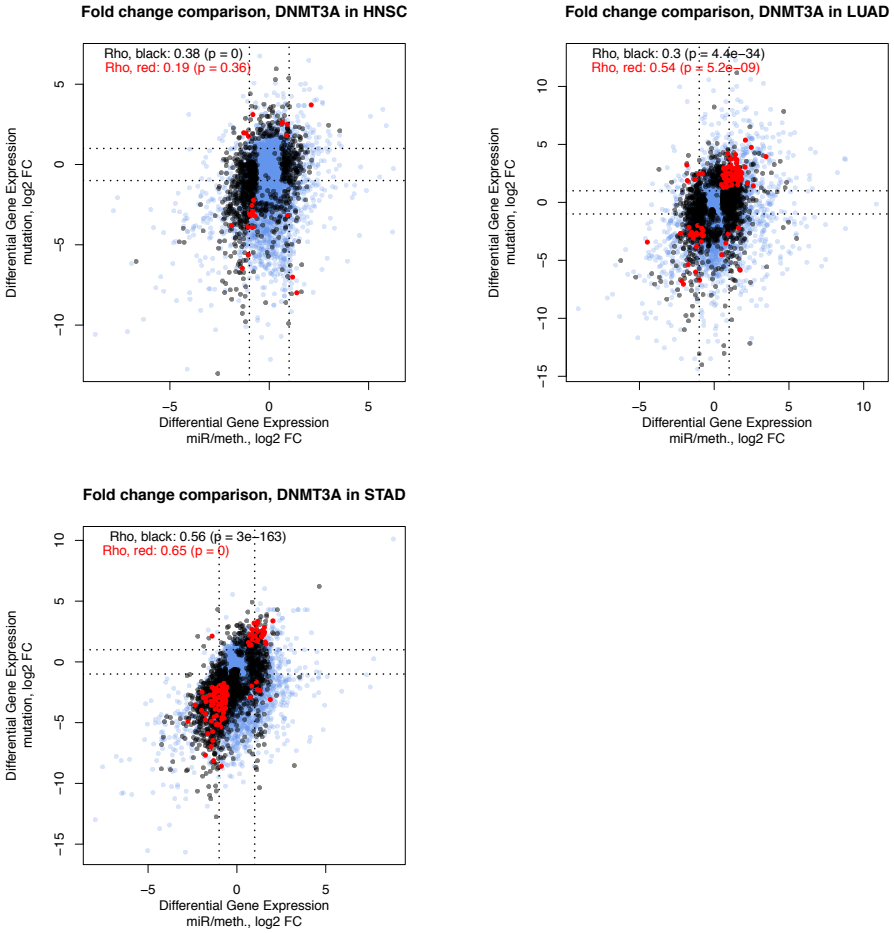

Figure B.33: **A similar profile of differential expression is observed in *DNMT3A* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *DNMT3A* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *DNMT3A* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

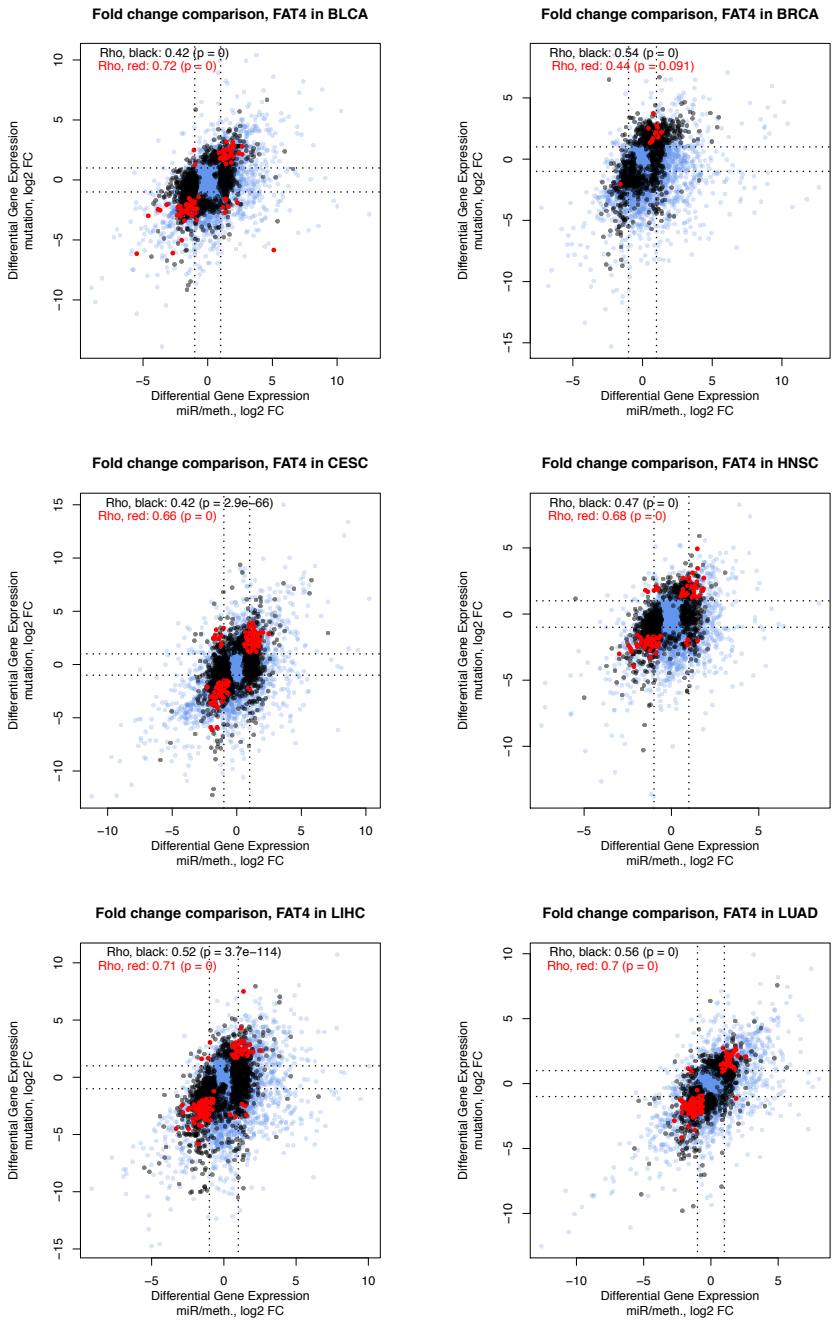

Figure B.34: **A similar profile of differential expression is observed in *FAT4* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *FAT4* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *FAT4* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

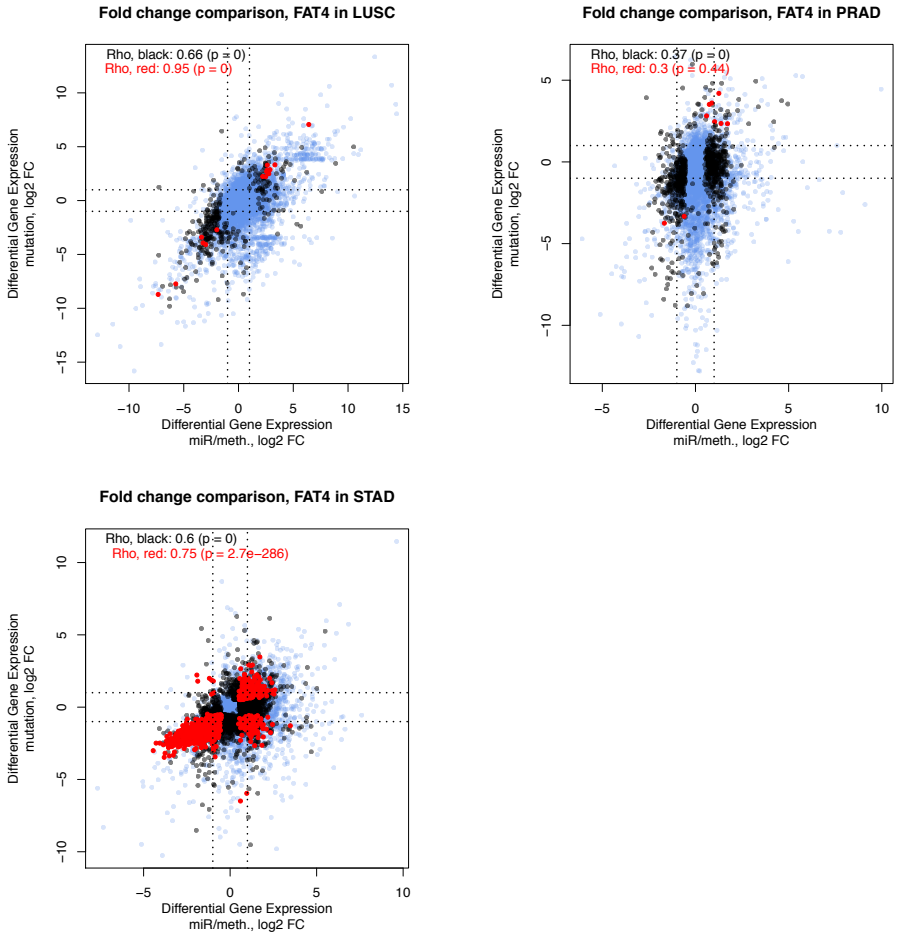

Figure B.35: **A similar profile of differential expression is observed in *FAT4* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *FAT4* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *FAT4* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

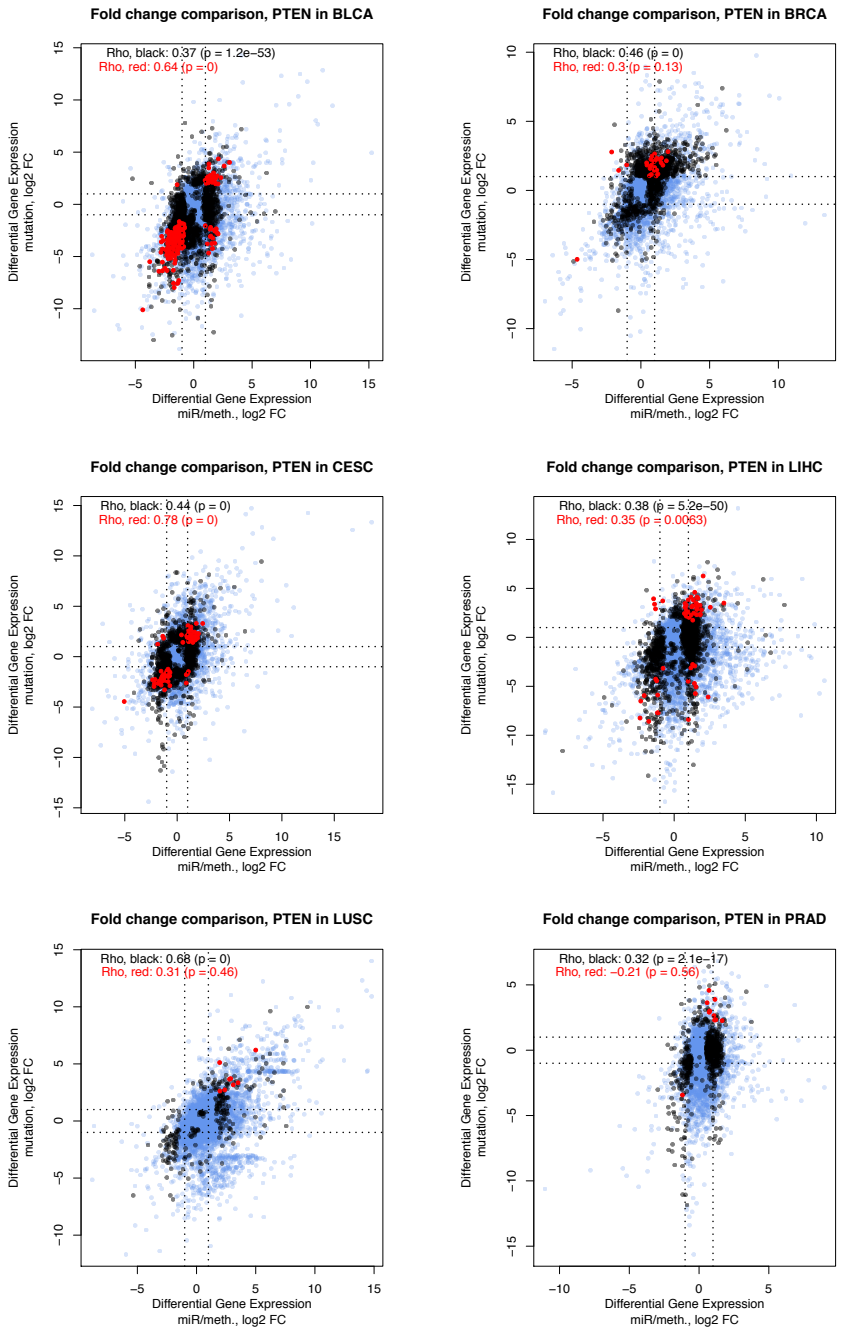

Figure B.36: **A similar profile of differential expression is observed in *PTEN* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *PTEN* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *PTEN* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

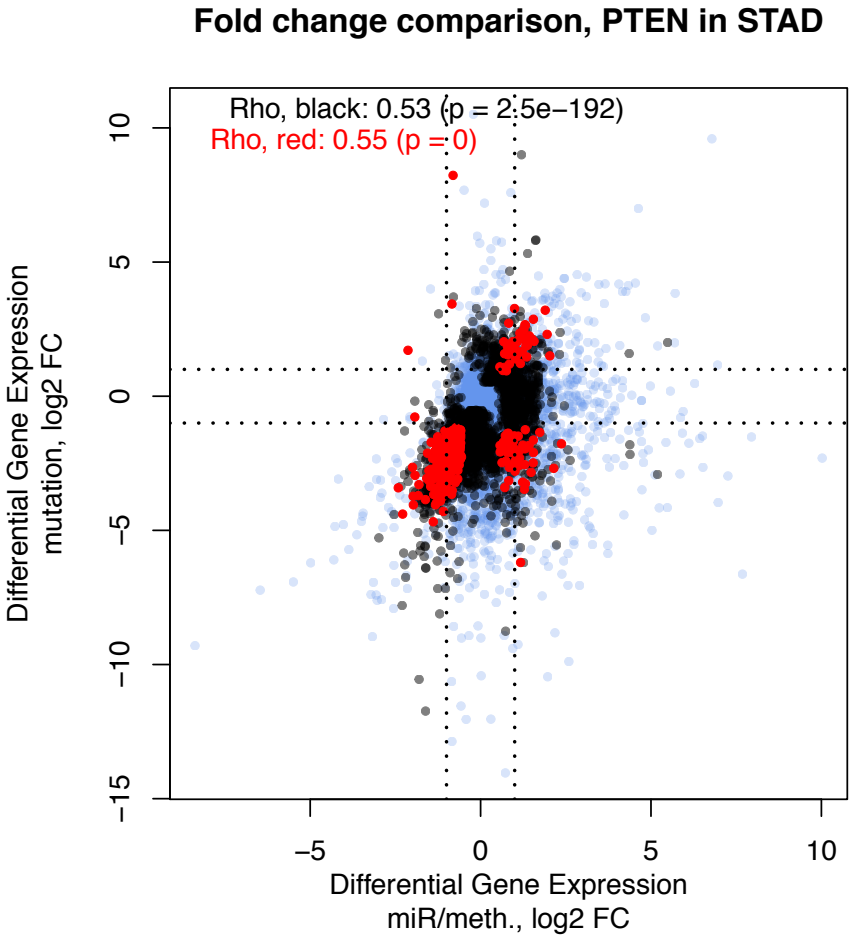

Figure B.37: **A similar profile of differential expression is observed in *PTEN* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *PTEN* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *PTEN* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

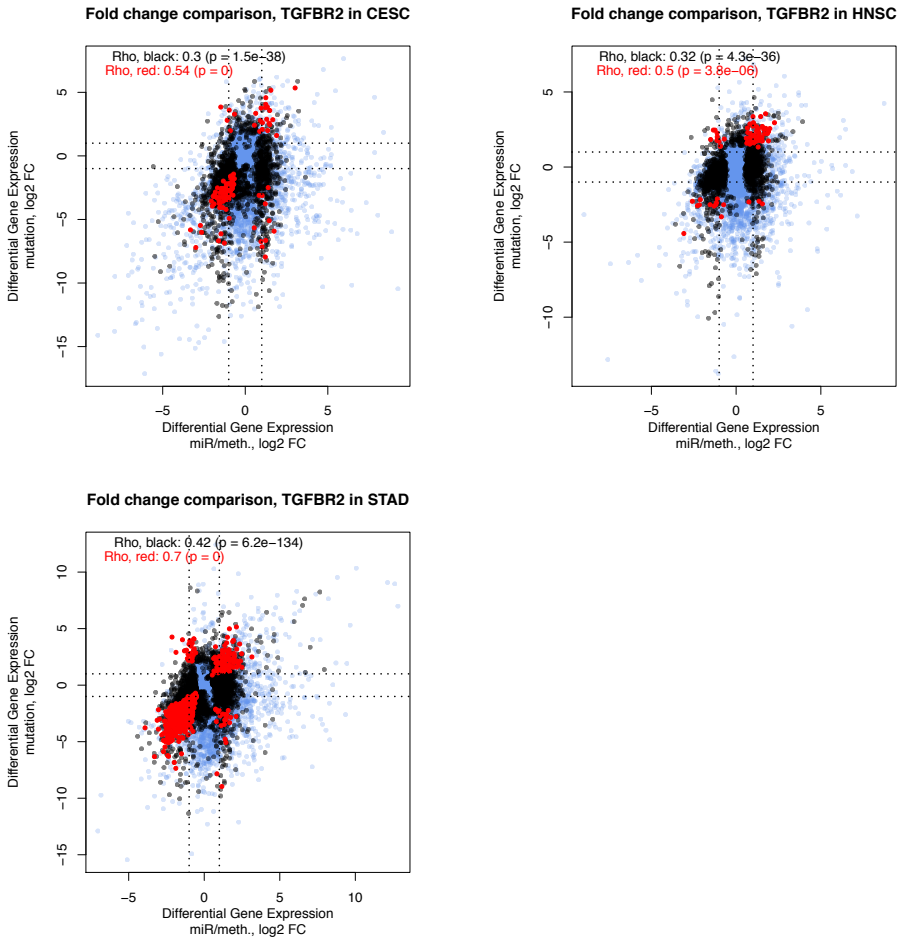

Figure B.38: **A similar profile of differential expression is observed in *TGFBR2* mutated cases as compared to miRNA/methylation high cases.** Fold change (log2 transformed) for differentially expressed genes in *TGFBR2* mutated cases versus non-mutated, miRNA low, methylation-low samples was plotted against fold change (log2 transformed) for differentially expressed genes in unmutated *TGFBR2* miRNA high and/or methylation high versus non-mutated, miRNA low, methylation-low samples. Genes in black are differentially expressed in one of the two groups, genes in red are commonly differentially expressed, and genes represented by blue points are not differentially expressed in either case. Spearman's rho is computed both for black points and red points as given in plots.

B.8 MYC amplification status and TSG-associated miRNA expression

Figure B.39: **miRNA expression largely does not show association with *MYC* amplification status across cancer types.** Heatmap of the p values obtained for the Wilcoxon rank-sum test, one-sided, comparing TSG-associated miRNA expression for *MYC* amplified and non-amplified cases, across tumour types (alternative hypothesis miRNA expression greater in amplified cases). NA values are indicated in grey.

Bibliography

[1] Roy S Herbst. Review of epidermal growth factor receptor biology. *International Journal of Radiation Oncology Biology Physics*, 59(2):S21–S26, 2004.

[2] Elizabeth Yang, Jiping Zha, Jennifer Jockel, Lawrence H Boise, Craig B Thompson, and Stanley J Korsmeyer. Bad, a heterodimeric partner for Bcl-XL and Bcl-2, displaces Bax and promotes cell death. *Cell*, 80(2):285–291, 1995.

[3] Jacob Scott and Andriy Marusyk. Somatic clonal evolution: a selection-centric perspective. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1867(2):139–150, 2017.

[4] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[5] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[6] Yuri Lazebnik. What are the hallmarks of cancer? *Nature Reviews Cancer*, 10(4):232, 2010.

[7] Carlos Sonnenschein and Ana M Soto. The aging of the 2000 and 2011 Hallmarks of Cancer reviews: a critique. *Journal of Biosciences*, 38(3):651–663, 2013.

[8] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

[9] Gene-Wei Li and X Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308, 2011.

[10] Anna Bergamaschi, Young H Kim, Pei Wang, Therese Sørlie, Tina Hernandez-Boussard, Per E Lonning, Robert Tibshirani, Anne-Lise Børresen-Dale, and Jonathan R Pollack. Distinct patterns of DNA copy number alteration are

associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer*, 45(11):1033–1040, 2006.

[11] Stephen B Baylin. DNA methylation and gene silencing in cancer. *Nature Reviews Clinical Oncology*, 2(S1):S4, 2005.

[12] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[13] Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, 27(1):15–26, 2015.

[14] Elza C de Bruin, Nicholas McGranahan, Richard Mitter, Max Salm, David C Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J Rowan, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206):251–256, 2014.

[15] Alexander RA Anderson, Alissa M Weaver, Peter T Cummings, and Vito Quaranta. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell*, 127(5):905–915, 2006.

[16] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338, 2013.

[17] Marco Gerlinger, Stuart Horswell, James Larkin, Andrew J Rowan, Max P Salm, Ignacio Varela, Rosalie Fisher, Nicholas McGranahan, Nicholas Matthews, Claudio R Santos, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3):225, 2014.

[18] Andrew Dhawan, Trevor A Graham, and Alexander G Fletcher. A computational modeling approach for deriving biomarkers to predict cancer risk in premalignant disease. *Cancer Prevention Research*, 9(4):283–295, 2016.

[19] Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non-small cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[20] Adam Humphries, Biancastella Cereser, Laura J Gay, Daniel SJ Miller, Bibek Das, Alice Gutteridge, George Elia, Emma Nye, Rosemary Jeffery, Richard Poulsom, et al. Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proceedings of the National Academy of Sciences*, 110(27):E2490–E2499, 2013.

[21] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart, and Christos Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, 14(16):5158–5165, 2008.

[22] Roman Rouzier, Charles M Perou, W Fraser Symmans, Nuhad Ibrahim, Massimo Cristofanilli, Keith Anderson, Kenneth R Hess, James Stec, Mark Ayers, Peter Wagner, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical Cancer Research*, 11(16):5678–5685, 2005.

[23] Joshua D Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, page eaar3247, 2018.

[24] Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. *Broad Institute TCGA Genome Data Analysis Center*.

[25] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[26] Alexander F Palazzo and Eliza S Lee. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, 6:2, 2015.

[27] John S Mattick and Igor V Makunin. Non-coding RNA. *Human Molecular Genetics*, 15(suppl_1):R17–R29, 2006.

[28] Tony Gutschner and Sven Diederichs. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biology*, 9(6):703–719, 2012.

[29] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D Mackowiak, Lea H Gregersen, Mathias Munschauer, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333, 2013.

[30] ENCODE Project Consortium et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[31] Eric M Small and Eric N Olson. Pervasive roles of microRNAs in cardiovascular biology. *Nature*, 469(7330):336, 2011.

[32] S Yendamuri and GA Calin. The role of microRNA in human leukemia: a review. *Leukemia*, 23(7):1257, 2009.

[33] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microRNA–target recognition. *PLoS Biology*, 3(3):e85, 2005.

[34] Hiro-oki Iwakawa and Yukihide Tomari. The functions of microRNAs: mRNA decay and translational repression. *Trends in Cell Biology*, 25(11):651–665, 2015.

[35] Leigh-Ann MacFarlane and Paul R Murphy. MicroRNA: biogenesis, function and role in cancer. *Current Genomics*, 11(7):537–561, 2010.

[36] Joerg E Braun, Eric Huntzinger, Maria Fauser, and Elisa Izaurralde. GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. *Molecular Cell*, 44(1):120–133, 2011.

[37] Hansruedi Mathys, Jérôme Basquin, Sevim Ozgur, Mariusz Czarnocki-Cieciura, Fabien Bonneau, Aafke Aartse, Andrzej Dziembowski, Marcin Nowotny, Elena Conti, and Witold Filipowicz. Structural and biochemical insights to the role of the CCR4-NOT complex and DDX6 ATPase in microRNA repression. *Molecular Cell*, 54(5):751–765, 2014.

[38] Géraldine Mathonnet, Marc R Fabian, Yuri V Svitkin, Armen Parsyan, Laurent Huck, Takayuki Murata, Stefano Biffo, William C Merrick, Edward Darzynkiewicz, Ramesh S Pillai, et al. MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science*, 317(5845):1764–1767, 2007.

[39] Shuibin Lin and Richard I Gregory. MicroRNA biogenesis pathways in cancer. *Nature Reviews Cancer*, 15(6):321, 2015.

[40] Shobha Vasudevan, Yingchun Tong, and Joan A Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–1934, 2007.

[41] Shobha Vasudevan and Joan A Steitz. AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell*, 128(6):1105–1118, 2007.

[42] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, 23(20):4051–4060, 2004.

[43] Antony Rodriguez, Sam Griffiths-Jones, Jennifer L Ashurst, and Allan Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Research*, 14(10a):1902–1910, 2004.

[44] David P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[45] Yoontae Lee, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim, and V Narry Kim. MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, 21(17):4663–4670, 2002.

[46] Markus T Bohnsack, Kevin Czaplinski, and DIRK GÖRLICH. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 10(2):185–191, 2004.

[47] Sven Diederichs and Daniel A Haber. Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression. *Cell*, 131(6):1097–1108, 2007.

[48] Haidi Zhang, Fabrice A Kolb, Vincent Brondani, Eric Billy, and Witold Filipowicz. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *The EMBO Journal*, 21(21):5875–5885, 2002.

[49] Tim A Rand, Sean Petersen, Fenghe Du, and Xiaodong Wang. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell*, 123(4):621–629, 2005.

[50] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[51] Steven R Head, H Kiyomi Komori, Sarah A LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56(2):61, 2014.

[52] Robert A Holt and Steven JM Jones. The new paradigm of flow cell sequencing. *Genome Research*, 18(6):839–846, 2008.

[53] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135, 2008.

[54] I Illumina. An introduction to next-generation sequencing technology. 2015.

[55] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[56] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[57] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[58] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, 2013.

[59] Martin Kircher, Patricia Heyn, and Janet Kelso. Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics*, 12(1):382, 2011.

[60] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36, 2012.

[61] Zhaojie Zhang, Jerome E Lee, Kent Riemondy, Emily M Anderson, and Rui Yi. High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biology*, 14(10):R109, 2013.

[62] Sam Griffiths-Jones, Russell J Grocock, Stijn Van Dongen, Alex Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl_1):D140–D144, 2006.

[63] Gert Van Peer, Steve Lefever, Jasper Anckaert, Anneleen Beckers, Ali Rihani, Alan Van Goethem, Pieter-Jan Volders, Fjoralba Zeka, Maté Ongenaert, Pieter Mestdagh, et al. miRBase Tracker: keeping track of microRNA annotation changes. *Database*, 2014:bau080, 2014.

[64] Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[65] Rui Liu, Xinhao Wang, Grace Y Chen, Piero Dalerba, Austin Gurney, Timothy Hoey, Gavin Sherlock, John Lewicki, Kerby Shedden, and Michael F Clarke. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New England Journal of Medicine*, 356(3):217–226, 2007.

[66] Lauren Averett Byers, Lixia Diao, Jing Wang, Pierre Saintigny, Luc Girard, Michael Peyton, Li Shen, Youhong Fan, Uma Giri, Praveen K Tumula, et al. An epithelial–mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical Cancer Research*, 19(1):279–290, 2013.

[67] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004.

[68] Elzbieta A Slodkowska and Jeffrey S Ross. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Review of Molecular Diagnostics*, 9(5):417–422, 2009.

[69] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[70] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3(1):246–259, 2013.

[71] Rashmi Kanagal-Shamanna, Bryce P Portier, Rajesh R Singh, Mark J Routbort, Kenneth D Aldape, Brian A Handal, Hamed Rahimi, Neelima G Reddy, Bedia A Barkoh, Bal M Mishra, et al. Next-generation sequencing-based multigene mutation profiling of solid tumors using fine needle aspiration samples: promises and challenges for routine clinical diagnostics. *Modern Pathology*, 27(2):314, 2014.

[72] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[73] Aiguo Li, Jennifer Walling, Susie Ahn, Yuri Kotliarov, Qin Su, Martha Quezado, J Carl Oberholtzer, John Park, Jean C Zenklusen, and Howard A Fine. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Research*, 69(5):2091–2099, 2009.

[74] FM Buffa, AL Harris, CM West, and CJ Miller. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British Journal of Cancer*, 102(2):428–435, 2010.

[75] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[76] David Venet, Jacques E Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240, 2011.

[77] Anders E Berglund, Eric A Welsh, and Steven A Eschrich. Characteristics and Validation Techniques for PCA-Based Gene-Expression Signatures. *International Journal of Genomics*, 2017, 2017.

[78] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

[79] Almut Schulze and Julian Downward. Navigating gene expression using microarrays - a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001.

[80] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one*, 9(1):e78644, 2014.

[81] Charles Wang, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, Huixiao Hong, Jie Shen, Zhenqiang Su, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology*, 32(9):926, 2014.

[82] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

[83] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[84] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

[85] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[86] A Subramanian, P Tamayo, VK Mootha, Sayan Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, and JP Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[87] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14(1):7, 2013.

[88] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108, 2009.

[89] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):225, 2005.

[90] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, pages 107–129, 2007.

[91] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68, 2015.

[92] Lance D Miller, Johanna Smeds, Joshy George, Vinsensius B Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T Liu, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, 102(38):13550–13555, 2005.

[93] Natalie S Fox, Maud HW Starmans, Syed Haider, Philippe Lambin, and Paul C Boutros. Ensemble analyses improve signatures of tumour hypoxia and reveal inter-platform differences. *BMC Bioinformatics*, 15(1):170, 2014.

[94] Steen Knudsen, Thomas Jensen, Anker Hansen, Wiktor Mazin, Justin Lindemann, Irene Kuter, Naomi Laing, and Elizabeth Anderson. Development and validation of a gene expression score that predicts response to fulvestrant in breast cancer patients. *PLoS One*, 9(2):e87415, 2014.

[95] Hung-I Harry Chen, Tzu-Hung Hsiao, Yidong Chen, and Charles Keller. S-score: A novel scoring method of gene signatures for molecular classification. In *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on*, pages 154–157. IEEE, 2011.

[96] Tzu-Hung Hsiao, Hung-I Harry Chen, Jo-Yang Lu, Pei-Ying Lin, Charles Keller, Sarah Comerford, Gail E Tomlinson, and Yidong Chen. Utilizing signature-score to identify oncogenic pathways of cholangiocarcinoma. *Translational Cancer Research*, 2(1):6, 2013.

[97] Hiromichi Ebi, Shuta Tomida, Toshiyuki Takeuchi, Chinatsu Arima, Takahiko Sato, Tetsuya Mitsudomi, Yasushi Yatabe, Hirotaka Osada, and Takashi Takahashi. Relationship of deregulated signaling converging onto mTOR with prognosis and classification of lung adenocarcinoma shown by two independent in silico analyses. *Cancer Research*, 69(9):4027–4035, 2009.

[98] Don L Gibbons, Wei Lin, Chad J Creighton, Shuling Zheng, Dror Berel, Yanan Yang, Maria Gabriela Raso, Diane D Liu, Ignacio I Wistuba, Guillermina Lozano, et al. Expression signatures of metastatic capacity in a genetic mouse model of lung adenocarcinoma. *PloS One*, 4(4):e5401, 2009.

[99] Sun Tian, Paul Roepman, Laura J Van't Veer, Rene Bernards, Femke De Snoo, and Annuska M Glas. Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. *Biomarker Insights*, 5:129, 2010.

[100] Anna V Roschke, Oleg K Glebov, Samir Lababidi, Kristen S Gehlhaus, John N Weinstein, and Ilan R Kirsch. Chromosomal instability is associated with higher expression of genes implicated in epithelial-mesenchymal transition, cancer invasiveness, and metastasis and with lower expression of genes involved in cell cycle checkpoints, DNA repair, and chromatin maintenance. *Neoplasia*, 10(11):1222IN10–1230IN26, 2008.

[101] Jon Jones, Hasan Otu, Dimitrios Spentzos, Shakirahmed Kolia, Mehmet Inan, Wolf D Beecken, Christian Fellbaum, Xuesong Gu, Marie Joseph, Allan J Pantuck, et al. Gene signatures of progression and metastasis in renal cell cancer. *Clinical Cancer Research*, 11(16):5730–5739, 2005.

[102] Andrea H Bild, Guang Yao, Jeffrey T Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M Lancaster, Andrew Berchuck, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353, 2006.

[103] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics*, 16(7):421–433, 2015.

[104] Gabriel B Loeb, Aly A Khan, David Canner, Joseph B Hiatt, Jay Shendure, Robert B Darnell, Christina S Leslie, and Alexander Y Rudensky. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Molecular Cell*, 48(5):760–770, 2012.

[105] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microRNA targets predicts functional nonconserved and non-canonical sites. *Genome Biology*, 11(8):R90, 2010.

[106] Ross Cloney. Non-coding RNA: Deciphering the rules of microRNA targeting. *Nature Reviews Genetics*, 17(12):718, 2016.

[107] Jean Hausser and Mihaela Zavolan. Identification and consequences of miRNA–target interactions?beyond repression of gene expression. *Nature Reviews Genetics*, 15(9):599–612, 2014.

[108] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:e05005.

[109] Cameron P Bracken, Hamish S Scott, and Gregory J Goodall. A network-biology perspective of microRNA function and dysfunction in cancer. *Nature Reviews Genetics*, 17(12):719–732, 2016.

[110] Francesca M Buffa, Carme Camps, Laura Winchester, Cameron E Snell, Harriet E Gee, Helen Sheldon, Marian Taylor, Adrian L Harris, and Jiannis Ragoussis. microRNA associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Research*, pages canres–0489, 2011.

[111] Heidi Dvinge, Anna Git, Stefan Gräf, Mali Salmon-Divon, Christina Curtis, Andrea Sottoriva, Yongjun Zhao, Martin Hirst, Javier Armisen, Eric A Miska, et al. The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature*, 497(7449):378–382, 2013.

[112] Cameron W Brennan, Roel GW Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty,

J Zachary Sanborn, Samuel H Berman, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, 2013.

[113] Caleb F Davis, Christopher J Ricketts, Min Wang, Lixing Yang, Andrew D Cherniack, Hui Shen, Christian Buhay, Hyojin Kang, Sang Cheol Kim, Catherine C Fahey, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3):319–330, 2014.

[114] Hamid Bolouri, Jason E Farrar, Timothy Triche Jr, Rhonda E Ries, Emilia L Lim, Todd A Alonzo, Yussanne Ma, Richard Moore, Andrew J Mungall, Marco A Marra, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nature Medicine*, 24(1):103–112, 2018.

[115] Anders Jacobsen, Joachim Silber, Girish Harinath, Jason T Huse, Nikolaus Schultz, and Chris Sander. Analysis of microRNA-target interactions across diverse cancer types. *Nature Structural and Molecular Biology*, 20(11):1325–1332, 2013.

[116] Yuqing Zhang, Min Li, Hao Wang, William E Fisher, Peter H Lin, Qizhi Yao, and Changyi Chen. Profiling of 95 micrornas in pancreatic cancer cell lines and surgical specimens by real-time pcr analysis. *World journal of surgery*, 33(4):698, 2009.

[117] Li-Fan Lu, Georg Gasteiger, I-Shing Yu, Ashutosh Chaudhry, Jing-Ping Hsin, Yuheng Lu, Paula D Bos, Ling-Li Lin, Carolyn L Zawislak, Sunglim Cho, et al. A single miRNA-mRNA interaction affects the immune response in a context- and cell-type-specific manner. *Immunity*, 43(1):52–64, 2015.

[118] Xiaohong Wang, Xiaowei Zhang, Xiao-Ping Ren, Jing Chen, Hongzhu Liu, Junqi Yang, Mario Medvedovic, Zhuowei Hu, and Guo-Chang Fan. MicroRNA-494 targeting both pro-apoptotic and anti-apoptotic proteins protects against ischemia/reperfusion-induced cardiac injury. *Circulation*, 122(13):1308, 2010.

[119] Adam P Carroll, Paul A Tooney, and Murray J Cairns. Context-specific microRNA function in developmental complexity. *Journal of Molecular Cell Biology*, 5(2):73–84, 2013.

[120] Fen-Biao Gao. Context-dependent functions of specific microRNAs in neuronal development. *Neural Development*, 5(1):25, 2010.

[121] Rémy Denzler, Sean E McGeary, Alexandra C Title, Vikram Agarwal, David P Bartel, and Markus Stoffel. Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Molecular Cell*, 64(3):565–579, 2016.

[122] Hua-Sheng Chiu, María Rodríguez Martínez, Mukesh Bansal, Aravind Subramanian, Todd R Golub, Xuerui Yang, Pavel Sumazin, and Andrea Califano. High-throughput validation of ceRNA regulatory networks. *BMC Genomics*, 18(1):418, 2017.

[123] Juan Xu, Yongsheng Li, Jianping Lu, Tao Pan, Na Ding, Zishan Wang, Tingting Shao, Jinwen Zhang, Lihua Wang, and Xia Li. The mRNA related ceRNA–ceRNA landscape and significance across 20 major cancer types. *Nucleic Acids Research*, 43(17):8169–8182, 2015.

[124] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[125] Junjun Zhang, Joachim Baran, Anthony Cros, Jonathan M Guberman, Syed Haider, Jack Hsu, Yong Liang, Elena Rivkin, Jianxin Wang, Brett Whitty, et al. International Cancer Genome Consortium Data Portal - a one-stop shop for cancer genomics data. *Database*, 2011, 2011.

[126] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[127] Gabor Csardi. *targetscan.Hs.eg.db: TargetScan miRNA target predictions for human*, 2013. R package version 0.6.1.

[128] Melike Marsan, Gert Van den Eynden, Ridha Limame, Patrick Neven, Jan Hauspy, Peter A Van Dam, Ignace Vergote, Luc Y Dirix, Peter B Vermeulen, and Steven J Van Laere. A core invasiveness gene signature reflects epithelial-to-mesenchymal transition but not metastatic potential in breast cancer cell lines and tissue samples. *PloS One*, 9(2):e89262, 2014.

[129] Massimo Masiero, Filipa Costa Simões, Hee Dong Han, Cameron Snell, Tessa Peterkin, Esther Bridges, Lingegowda S Mangala, Sherry Yen-Yao Wu, Sunila Pradeep, Demin Li, et al. A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor ELTD1 as a key regulator of angiogenesis. *Cancer Cell*, 24(2):229–241, 2013.

[130] Jelle J Goeman. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1):70–84, 2010.

[131] Jelle J. Goeman. *Penalized R package, version 0.9-50*, 2017. R package.

[132] Francesco Del Carratore, Andris Janckevics Fangxin Hong <fx-hong@jimmy.harvard.edu>, Ben Wittner, Rainer Breitling, , and Florian Battke. *RankProd: Rank Product method for identifying differentially expressed genes with application in meta-analysis*, 2016. R package version 3.0.0.

[133] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1-3):83–92, 2004.

[134] Maciej Pajak and T. Ian Simpson. *miRNAtap: miRNAtap: microRNA Targets - Aggregated Predictions*, 2016. R package version 1.8.0.

[135] Manolis Maragkakis, Martin Reczko, Victor A Simossis, Panagiotis Alexiou, Giorgos L Papadopoulos, Theodore Dalamagas, Giorgos Giannopoulos, G Goumas, Evangelos Koukis, Kornilios Kourtis, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research*, 37(suppl_2):W273–W276, 2009.

[136] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

[137] Sabbi Lall, Dominic Grün, Azra Krek, Kevin Chen, Yi-Lu Wang, Colin N Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip MacMenamin, et al. A genome-wide map of conserved microRNA targets in C. elegans. *Current Biology*, 16(5):460–471, 2006.

[138] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[139] Nathan Wong and Xiaowei Wang. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Research*, 43(D1):D146–D152, 2014.

[140] Seongho Kim. *ppcor: Partial and Semi-Partial (Part) Correlation*, 2015. R package version 1.1.

[141] Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2016.

[142] Andrew Dhawan, Alessandro Barberis, Wei-Chen Cheng, Enric Domingo, Catharine West, Tim Maughan, Jacob Scott, Adrian L Harris, and Francesca M Buffa. sigQC: A procedural approach for standardising the evaluation of gene signatures. *bioRxiv*, page 203729, 2017.

[143] Gregg L Semenza. HIF-1: mediator of physiological and pathophysiological responses to hypoxia. *Journal of Applied Physiology*, 88(4):1474–1480, 2000.

[144] Carme Camps, Francesca M Buffa, Stefano Colella, John Moore, Christos Sotiriou, Helen Sheldon, Adrian L Harris, Jonathan M Gleadle, and Jiannis Ragoussis. hsa-miR-210 Is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clinical Cancer Research*, 14(5):1340–1348, 2008.

[145] Ritu Kulshreshtha, Manuela Ferracin, Sylwia E Wojcik, Ramiro Garzon, Hansjuerg Alder, Francisco J Agosto-Perez, Ramana Davuluri, Chang-Gong Liu, Carlo M Croce, Massimo Negrini, et al. A microRNA signature of hypoxia. *Molecular and Cellular Biology*, 27(5):1859–1867, 2007.

[146] Zhen Chen, Tsung-Ching Lai, Yi-Hua Jan, Feng-Mao Lin, Wei-Chi Wang, Han Xiao, Yun-Ting Wang, Wei Sun, Xiaopei Cui, Ying-Shiuan Li, et al. Hypoxia-responsive miRNAs target argonaute 1 to promote angiogenesis. *The Journal of Clinical Investigation*, 123(3):1057, 2013.

[147] Munekazu Yamakuchi, Shusuke Yagi, Takashi Ito, and Charles J Lowenstein. MicroRNA-22 regulates hypoxia signaling in colon cancer cells. *PloS One*, 6(5):e20291, 2011.

[148] Timothy KK Kamanu, Aleksandar Radovanovic, John AC Archer, and Vladimir B Bajic. Exploration of miRNA families for hypotheses generation. *Scientific Reports*, 3, 2013.

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

[150] Maciej Pajak and T Ian Simpson. miRNAtap. db: microRNA Targets-Aggregated Predictions database use. 2014.

[151] Sandrine Tchatchou, Anke Jung, Kari Hemminki, Christian Sutter, Barbara Wappenschmidt, Peter Bugert, Bernhard HF Weber, Dieter Niederacher, Norbert Arnold, Raymonda Varon-Mateeva, et al. A variant affecting a putative miRNA target site in estrogen receptor (ESR) 1 is associated with breast cancer risk in premenopausal women. *Carcinogenesis*, 30(1):59–64, 2008.

[152] Wan-Hsin Liu, Shiou-Hwei Yeh, Cho-Chun Lu, Sung-Liang Yu, Hsuan-Yu Chen, Chien-Yu Lin, Ding-Shinn Chen, and Pei-Jer Chen. MicroRNA-18a prevents estrogen receptor-α expression, promoting proliferation of hepatocellular carcinoma cells. *Gastroenterology*, 136(2):683–693, 2009.

[153] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[154] Dong Wang, Juan Wang, Ming Lu, Fei Song, and Qinghua Cui. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, 26(13):1644–1650, 2010.

[155] Ming Lu, Qipeng Zhang, Min Deng, Jing Miao, Yanhong Guo, Wei Gao, and Qinghua Cui. An analysis of human microRNA and disease associations. *PLOS One*, 3(10):e3420, 2008.

[156] Rémy Denzler, Vikram Agarwal, Joanna Stefano, David P Bartel, and Markus Stoffel. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Molecular Cell*, 54(5):766–776, 2014.

[157] Claire L Smillie, Tamara Sirey, and Chris P Ponting. Complexities of post-transcriptional regulation and the modeling of ceRNA crosstalk. *Critical Reviews in Biochemistry and Molecular Biology*, 53(3):231–245, 2018.

[158] Kourosh Zarringhalam, Yvonne Tay, Prajna Kulkarni, Assaf C Bester, Pier Paolo Pandolfi, and Rahul V Kulkarni. Identification of competing endogenous RNAs of the tumor suppressor gene PTEN: A probabilistic approach. *Scientific Reports*, 7(1):7755, 2017.

[159] Emanuele de Rinaldis, Patrycja Gazinska, Anca Mera, Zora Modrusan, Grazyna M Fedorowicz, Brian Burford, Cheryl Gillett, Pierfrancesco Marra, Anita Grigoriadis, David Dornan, et al. Integrated genomic analysis of triple-negative breast cancers reveals novel microRNAs associated with clinical and molecular phenotypes and sheds light on the pathways they control. *BMC Genomics*, 14(1):643, 2013.

[160] Ting Gong and Joseph D Szustakowski. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, 29(8):1083–1085, 2013.

[161] Andrew J Gentles, Aaron M Newman, Chih Long Liu, Scott V Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S Nair, Yue Xu, Amanda Khuong, Chuong D Hoang, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature Medicine*, 21(8):938–945, 2015.

[162] Yanjie Lu, Jiening Xiao, Huixian Lin, Yunlong Bai, Xiaobin Luo, Zhiguo Wang, and Baofeng Yang. A single anti-microRNA antisense oligodeoxyribonucleotide (AMO) targeting multiple microRNAs offers an improved approach for microRNA interference. *Nucleic Acids Research*, 37(3):e24–e24, 2009.

[163] Yijun Yang, Fei Li, Manujendra N Saha, Jahangir Abdi, Lugui Qiu, and Hong Chang. miR-137 and miR-197 induce apoptosis and suppress tumorigenicity by targeting MCL-1 in multiple myeloma. *Clinical Cancer Research*, 21(10):2399–2411, 2015.

[164] Sun Hee Lee, Yuk Dong Jung, Young Sun Choi, and You Mie Lee. Targeting of RUNX3 by miR-130a and miR-495 cooperatively increases cell proliferation and tumor angiogenesis in gastric cancer cells. *Oncotarget*, 6(32):33269, 2015.

[165] Eleonora Brognara, Enrica Fabbri, Giulia Montagner, Jessica Gasparello, Alex Manicardi, Roberto Corradini, Nicoletta Bianchi, Alessia Finotti, Giulia Breveglieri, Monica Borgatti, et al. High levels of apoptosis are induced in human glioma cell lines by co-administration of peptide nucleic acids targeting miR-221 and miR-222. *International Journal of Oncology*, 48(3):1029–1038, 2016.

[166] Chunzhi Zhang, Chunsheng Kang, Yongping You, Peiyu Pu, Weidong Yang, Peng Zhao, Guangxiu Wang, Anling Zhang, Zhifan Jia, Lei Han, et al. Co-suppression of miR-221/222 cluster suppresses human glioma cell growth by targeting p27kip1 in vitro and in vivo. *International Journal of Oncology*, 34(6):1653–1660, 2009.

[167] Bahar Yilmazel, Yanhui Hu, Frederic Sigoillot, Jennifer A Smith, Caroline E Shamu, Norbert Perrimon, and Stephanie E Mohr. Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis. *BMC Bioinformatics*, 15(1):192, 2014.

[168] Lars Aagaard and John J Rossi. RNAi therapeutics: principles, prospects and challenges. *Advanced Drug Delivery Reviews*, 59(2):75–86, 2007.

[169] Sacha I Rothschild. microRNA therapies in cancer. *Molecular and Cellular Therapies*, 2(1):7, 2014.

[170] Rajesha Rupaimoole and Frank J Slack. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nature Reviews Drug Discovery*, 16(3):203–222, 2017.

[171] Muhammad S Beg, Andrew J Brenner, Jasgit Sachdev, Mitesh Borad, Yoon-Koo Kang, Jay Stoudemire, Susan Smith, Andreas G Bader, Sinil Kim, and David S Hong. Phase I study of MRX34, a liposomal miR-34a mimic, administered twice weekly in patients with advanced solid tumors. *Investigational New Drugs*, 35(2):180–188, 2017.

[172] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[173] F Bertucci, P Finetti, and D Birnbaum. Basal breast cancer: a complex and deadly molecular subtype. *Current Molecular Medicine*, 12(1):96–110, 2012.

[174] Richard I Gregory and Ramin Shiekhattar. MicroRNA biogenesis and cancer. *Cancer Research*, 65(9):3509–3512, 2005.

[175] Tingfang Yi, Haribabu Arthanari, Barak Akabayov, Huaidong Song, Evangelos Papadopoulos, Hank H Qi, Mark Jedrychowski, Thomas Güttler, Cuicui Guo, Rafael E Luna, et al. eIF1A augments Ago2-mediated Dicer-independent miRNA biogenesis and RNA interference. *Nature Communications*, 6, 2015.

[176] Young-Kook Kim, Boseon Kim, and V Narry Kim. Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis. *Proceedings of the National Academy of Sciences*, 113(13):E1881–E1889, 2016.

[177] Minju Ha and V Narry Kim. Regulation of microrna biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8):509–524, 2014.

[178] Masaki Mori, Robinson Triboulet, Morvarid Mohseni, Karin Schlegelmilch, Kriti Shrestha, Fernando D Camargo, and Richard I Gregory. Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. *Cell*, 156(5):893–906, 2014.

[179] Rajesha Rupaimoole, Sherry Y Wu, Sunila Pradeep, Cristina Ivan, Chad V Pecot, Kshipra M Gharpure, Archana S Nagaraja, Guillermo N Armaiz-Pena, Michael McGuire, Behrouz Zand, et al. Hypoxia-mediated downregulation of miRNA biogenesis promotes tumour progression. *Nature Communications*, 5:5202, 2014.

[180] Twan Van Den Beucken, Elizabeth Koch, Kenneth Chu, Rajesha Rupaimoole, Peggy Prickaerts, Michiel Adriaens, Jan Willem Voncken, Adrian L Harris, Francesca M Buffa, Syed Haider, et al. Hypoxia promotes stem cell phenotypes and poor prognosis through epigenetic regulation of DICER. *Nature Communications*, 5:5203, 2014.

[181] Sidi Chen, Yuan Xue, Xuebing Wu, Cong Le, Arjun Bhutkar, Eric L Bell, Feng Zhang, Robert Langer, and Phillip A Sharp. Global microRNA depletion suppresses tumor angiogenesis. *Genes & Development*, 28(10):1054–1067, 2014.

[182] Weidong Yang, Thimmaiah P Chendrimada, Qingde Wang, Miyoko Higuchi, Peter H Seeburg, Ramin Shiekhattar, and Kazuko Nishikura. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nature Structural & Molecular Biology*, 13(1):13–21, 2006.

[183] Mathilde Fagard, Stéphanie Boutet, Jean-Benoit Morel, Catherine Bellini, and Hervé Vaucheret. AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proceedings of the National Academy of Sciences*, 97(21):11650–11654, 2000.

[184] George L Sen and Helen M Blau. Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. *Nature Cell Biology*, 7(6):633–636, 2005.

[185] Mayuko Yoda, Tomoko Kawamata, Zain Paroo, Xuecheng Ye, Shintaro Iwasaki, Qinghua Liu, and Yukihide Tomari. ATP-dependent human RISC assembly pathways. *Nature Structural & Molecular Biology*, 17(1):17–23, 2010.

[186] Leah R Sabin, Rui Zhou, Joshua J Gruber, Nina Lukinova, Shelly Bambina, Allison Berman, Chi-Kong Lau, Craig B Thompson, and Sara Cherry. Ars2 regulates both miRNA-and siRNA-dependent silencing and suppresses RNA virus infection in Drosophila. *Cell*, 138(2):340–351, 2009.

[187] Christopher Rouya, Nadeem Siddiqui, Masahiro Morita, Thomas F Duchaine, Marc R Fabian, and Nahum Sonenberg. Human DDX6 effects miRNA-mediated gene silencing via direct binding to CNOT1. *RNA*, 20(9):1398–1409, 2014.

[188] Jan Rehwinkel, Isabelle Behm-Ansmant, David Gatfield, and Elisa Izaurralde. A crucial role for GW182 and the DCP1: DCP2 decapping complex in miRNA-mediated gene silencing. *RNA*, 11(11):1640–1647, 2005.

[189] Hiroshi I Suzuki, Kaoru Yamagata, Koichi Sugimoto, Takashi Iwamoto, Shigeaki Kato, and Kohei Miyazono. Modulation of microRNA processing by p53. *Nature*, 460(7254):529–533, 2009.

[190] Brandi N Davis, Aaron C Hilyard, Giorgio Lagna, and Akiko Hata. SMAD proteins control DROSHA-mediated microRNA maturation. *Nature*, 454(7200):56–61, 2008.

[191] Akemi Takata, Motoyuki Otsuka, Takeshi Yoshikawa, Takahiro Kishikawa, Yotaro Kudo, Tadashi Goto, Haruhiko Yoshida, and Kazuhiko Koike. A miRNA machinery component DDX20 controls NF-κB via microRNA-140 function. *Biochemical and Biophysical Research Communications*, 420(3):564–569, 2012.

[192] Richard I Gregory, Kai-ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar. The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240, 2004.

[193] Teresa Lee and Jerry Pelletier. The biology of DHX9 and its potential as a therapeutic target. *Oncotarget*, 7(27):42716, 2016.

[194] Emily Bernstein, Amy A Caudy, Scott M Hammond, and Gregory J Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363, 2001.

[195] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rådmark, Sunyoung Kim, et al. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415, 2003.

[196] Nicole-Claudia Meisner and Witold Filipowicz. Properties of the regulatory RNA-binding protein HuR and its role in controlling miRNA repression. In *Regulation of microRNAs*, pages 106–123. Springer, 2010.

[197] Sandra L Romero-Cordoba, Ivan Salido-Guadarrama, Mauricio Rodriguez-Dorantes, and Alfredo Hidalgo-Miranda. miRNA biogenesis: biological impact in the development of cancer. *Cancer Biology & Therapy*, 15(11):1444–1455, 2014.

[198] Josee Dostie, Zissimos Mourelatos, Michael Yang, Anup Sharma, and Gideon Dreyfuss. Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA*, 9(2):180–186, 2003.

[199] Brandi N Davis and Akiko Hata. Regulation of MicroRNA Biogenesis: A miRiad of mechanisms. *Cell Communication and Signaling*, 7(1):18, 2009.

[200] Kiyoshi Masuda, Yuki Kuwano, Kensei Nishida, Kazuhito Rokutan, and Issei Imoto. NF90 in posttranscriptional gene regulation and microRNA biogenesis. *International Journal of Molecular Sciences*, 14(8):17111–17121, 2013.

[201] Yao Wei, Limin Li, Dong Wang, Chen-Yu Zhang, and Ke Zen. Importin 8 regulates the transport of mature microRNAs into the cell nucleus. *Journal of Biological Chemistry*, 289(15):10270–10275, 2014.

[202] Patrick Connerty, Sarah Bajan, Judit Remenyi, Frances V Fuller-Pace, and Gyorgy Hutvagner. The miRNA biogenesis factors, p72/DDX17 and KHSRP regulate the protein level of Ago2 in human cells. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(10):1299–1305, 2016.

[203] Shane R Horman, Maja M Janas, Claudia Litterst, Bingbing Wang, Ian J MacRae, Mary J Sever, David V Morrissey, Paul Graves, Biao Luo, Shaikamjad Umesalma, et al. Akt-mediated phosphorylation of argonaute 2 downregulates cleavage and upregulates translational repression of MicroRNA targets. *Molecular Cell*, 50(3):356–367, 2013.

[204] Phillip J Kenny, Hongjun Zhou, Miri Kim, Geena Skariah, Radhika S Khetani, Jenny Drnevich, Mary Luz Arcila, Kenneth S Kosik, and Stephanie Ceman. MOV10 and FMRP regulate AGO2 association with microRNA recognition elements. *Cell Reports*, 9(5):1729–1741, 2014.

[205] Robert W Walters, Shelton S Bradrick, and Matthias Gromeier. Poly (A)-binding protein modulates mRNA susceptibility to cap-dependent miRNA-mediated repression. *RNA*, 16(1):239–250, 2010.

[206] Shuai Li, Juanjuan Zhu, Hanjiang Fu, Jing Wan, Zheng Hu, Shanshan Liu, Jie Li, Yi Tie, Ruiyun Xing, Jie Zhu, et al. Hepato-specific microRNA-122 facilitates accumulation of newly synthesized miRNA through regulating PRKRA. *Nucleic Acids Research*, 40(2):884–891, 2011.

[207] Shelley Sazer and Mary Dasso. The ran decathlon: multiple roles of Ran. *J Cell Sci*, 113(7):1111–1118, 2000.

[208] Jung-Chun Lin and Woan-Yuh Tarn. RNA-binding motif protein 4 translocates to cytoplasmic granules and suppresses translation via argonaute2 during muscle cell differentiation. *Journal of Biological Chemistry*, 284(50):34658–34665, 2009.

[209] Reyad A Elbarbary, Keita Miyoshi, Omar Hedaya, Jason R Myers, and Lynne E Maquat. UPF1 helicase promotes TSN-mediated miRNA decay. *Genes & Development*, 31(14):1483–1493, 2017.

[210] Brandi N Davis, Aaron C Hilyard, Peter H Nguyen, Giorgio Lagna, and Akiko Hata. Smad proteins bind a conserved RNA sequence to promote microRNA maturation by Drosha. *Molecular Cell*, 39(3):373–384, 2010.

[211] Bin Yu, Liu Bi, Binglian Zheng, Lijuan Ji, David Chevalier, Manu Agarwal, Vanitharani Ramachandran, Wanxiang Li, Thierry Lagrange, John C Walker, et al. The FHA domain proteins DAWDLE in Arabidopsis and SNIP1 in humans act in small RNA biogenesis. *Proceedings of the National Academy of Sciences*, 105(29):10073–10078, 2008.

[212] Han Wu, Shuying Sun, Kang Tu, Yuan Gao, Bin Xie, Adrian R Krainer, and Jun Zhu. A splicing-independent function of SF2/ASF in microRNA processing. *Molecular Cell*, 38(1):67–77, 2010.

[213] Stasė Butkytė, Laurynas Čiupas, Eglė Jakubauskienė, Laurynas Vilys, Paulius Mocevicius, Arvydas Kanopka, and Giedrius Vilkaitis. Splicing-dependent expression of microRNAs of mirtron origin in human digestive and excretory system cancer cells. *Clinical Epigenetics*, 8(1):33, 2016.

[214] Thimmaiah P Chendrimada, Richard I Gregory, Easwari Kumaraswamy, Jessica Norman, Neil Cooch, Kazuko Nishikura, and Ramin Shiekhattar. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436(7051):740, 2005.

[215] Duygu Kuzuoğlu-Öztürk, Dipankar Bhandari, Eric Huntzinger, Maria Fauser, Sigrun Helms, and Elisa Izaurralde. miRISC and the CCR4–NOT complex silence mRNA targets independently of 43S ribosomal scanning. *The EMBO Journal*, 35(11):1186–1203, 2016.

[216] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics*, 11(9):597, 2010.

[217] Katsunori Aoki, Kenji Suzuki, Takashi Sugano, Tetsuya Tasaka, Kazuhiko Nakahara, Osamu Kuge, Akira Omori, and Masataka Kasai. A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nature Genetics*, 10(2):167, 1995.

[218] Kuo-Wang Tsai, Chung-Man Leung, Yi-Hao Lo, Ting-Wen Chen, Wen-Ching Chan, Shou-Yu Yu, Ya-Ting Tu, Hing-Chung Lam, Sung-Chou Li, Luo-Ping Ger, et al. Arm selection preference of microRNA-193a varies in breast cancer. *Scientific Reports*, 6:28176, 2016.

[219] Kimberly C Wiegand, Sohrab P Shah, Osama M Al-Agha, Yongjun Zhao, Kane Tse, Thomas Zeng, Janine Senz, Melissa K McConechy, Michael S Anglesio, Steve E Kalloger, et al. ARID1A mutations in endometriosis-associated ovarian carcinomas. *New England Journal of Medicine*, 363(16):1532–1543, 2010.

[220] SC Sun. CYLD: a tumor suppressor deubiquitinase regulating NF-κB activation and diverse biological processes. *Cell Death & Differentiation*, 17(1):25–34, 2010.

[221] Isabelle Sansal and William R Sellers. The biology and clinical relevance of the PTEN tumor suppressor pathway. *Journal of Clinical Oncology*, 22(14):2954–2963, 2004.

[222] Hao Huang, Xi Jiang, Zejuan Li, Yuanyuan Li, Chun-Xiao Song, Chunjiang He, Miao Sun, Ping Chen, Sandeep Gurbuxani, Jiapeng Wang, et al. TET1 plays an essential oncogenic role in MLL-rearranged leukemia. *Proceedings of the National Academy of Sciences*, 110(29):11994–11999, 2013.

[223] Jing Li, Ying Yang, Yue Peng, Richard J Austin, Winfried G van Eyndhoven, Ken CQ Nguyen, Tim Gabriele, Mila E McCurrach, Jeffrey R Marks, Timothy Hoey, et al. Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23. *Nature Genetics*, 31(2):133, 2002.

[224] CA French, CL Ramirez, J Kolmakova, TT Hickman, MJ Cameron, ME Thyne, JL Kutok, JA Toretsky, AK Tadavarthy, UR Kees, et al. BRD–NUT oncoproteins: a family of closely related nuclear proteins that block epithelial differentiation and maintain the growth of carcinoma cells. *Oncogene*, 27(15):2237–2242, 2008.

[225] Min S Kim, Ji E Oh, Yoo R Kim, Sang W Park, Mi R Kang, Sung S Kim, Chang H Ahn, Nam J Yoo, and Sug H Lee. Somatic mutations and losses of expression of microRNA regulation-related genes AGO2 and TNRC6A in gastric and colorectal cancers. *The Journal of Pathology*, 221(2):139–146, 2010.

[226] Jidong Liu, Fabiola V Rivas, James Wohlschlegel, John R Yates, Roy Parker, and Gregory J Hannon. A role for the P-body component GW182 in microRNA function. *Nature Cell Biology*, 7(12):1261–1266, 2005.

[227] Zhaohui Feng, Cen Zhang, Rui Wu, and Wenwei Hu. Tumor suppressor p53 meets microRNAs. *Journal of Molecular Cell Biology*, 3(1):44–50, 2011.

[228] Benjamin Boyerinas, Sun-Mi Park, Annika Hau, Andrea E Murmann, and Marcus E Peter. The role of let-7 in cell differentiation and cancer. *Endocrine-related Cancer*, 17(1):F19–F36, 2010.

[229] David W Salzman, Jonathan Shubert-Coleman, and Henry Furneaux. P68 RNA helicase unwinds the human let-7 microRNA precursor duplex and is required for let-7-directed silencing of gene expression. *Journal of Biological Chemistry*, 282(45):32773–32779, 2007.

[230] Stephen Y Chan, Ying-Yi Zhang, Craig Hemann, Christopher E Mahoney, Jay L Zweier, and Joseph Loscalzo. MicroRNA-210 controls mitochondrial metabolism during hypoxia by repressing the iron-sulfur cluster assembly proteins ISCU1/2. *Cell Metabolism*, 10(4):273–284, 2009.

[231] Min-Zu Wu, Wei-Chung Cheng, Su-Feng Chen, Shin Nieh, Carolyn O'Connor, Chia-Lin Liu, Wen-Wei Tsai, Cheng-Jang Wu, Lorena Martin, Yaoh-Shiang Lin, et al. miR-25/93 mediates hypoxia-induced immunosuppression by repressing cGAS. *Nature Cell Biology*, 19(10):1286, 2017.

[232] Daniel Triner and Yatrik M Shah. Hypoxia-inducible factors: a central link between inflammation and cancer. *The Journal of Clinical Investigation*, 126(10):3689–3698, 2016.

[233] Hedda A Meijer, Ewan M Smith, and Martin Bushell. Regulation of miRNA strand selection: follow the leader?, 2014.

[234] Wei-Ting Kuo, Ming-Wei Su, Yungling Leo Lee, Chien-Hsiun Chen, Chew-Wun Wu, Wen-Liang Fang, Kuo-Hung Huang, and Wen-chang Lin. Bioinformatic interrogation of 5p-arm and 3p-arm specific miRNA expression using TCGA datasets. *Journal of Clinical Medicine*, 4(9):1798–1814, 2015.

[235] Sung-Chou Li, Kuo-Wang Tsai, Hung-Wei Pan, Yung-Ming Jeng, Meng-Ru Ho, and Wen-Hsiung Li. MicroRNA 3'end nucleotide modification patterns and arm selection preference in liver tissues. *BMC Systems Biology*, 6(2):S14, 2012.

[236] María Tomé, Pedro López-Romero, Carmen Albo, Juan Carlos Sepúlveda, Benjamín Fernández-Gutiérrez, Ana Dopazo, Antonio Bernad, and Manuel A González. miR-335 orchestrates cell proliferation, migration and differentiation in human mesenchymal stem cells. *Cell Death & Differentiation*, 18(6):985–995, 2011.

[237] Zhi Yan, Yimin Xiong, Weitian Xu, Juan Gao, Yi Cheng, Zhigang Wang, Fang Chen, and Guorong Zheng. Identification of hsa-miR-335 as a prognostic signature in gastric cancer. *PloS One*, 7(7):e40037, 2012.

[238] Holger Heyn, Maria Engelmann, Sabine Schreek, Philipp Ahrens, Ulrich Lehmann, Hans Kreipe, Brigitte Schlegelberger, and Carmela Beger. MicroRNA miR-335 is crucial for the BRCA1 regulatory cascade in breast cancer development. *International Journal of Cancer*, 129(12):2797–2806, 2011.

[239] Minfeng Shu, Xiaoke Zheng, Sihan Wu, Huimin Lu, Tiandong Leng, Wenbo Zhu, Yuehan Zhou, Yanqiu Ou, Xi Lin, Yuan Lin, et al. Targeting oncogenic miR-335 inhibits growth and invasion of malignant astrocytoma cells. *Molecular Cancer*, 10(1):59, 2011.

[240] Cillian Clancy, Myles R Joyce, and Michael J Kerin. The use of circulating microRNAs as diagnostic biomarkers in colorectal cancer. *Cancer Biomarkers*, 15(2):103–113, 2015.

[241] Hai-Tao Zhu, Qiong-Zhu Dong, Yuan-Yuan Sheng, Jin-Wang Wei, Guan Wang, Hai-Jun Zhou, Ning Ren, Hu-Liang Jia, Qing-Hai Ye, and Lun-Xiu Qin. MicroRNA-29a-5p is a novel predictor for early recurrence of hepatitis B virus-related hepatocellular carcinoma after surgical resection. *PloS One*, 7(12):e52393, 2012.

[242] Zhujiang Zhao, Ling Wang, Wei Song, He Cui, Gang Chen, Fengchang Qiao, Jiaojiao Hu, Rongping Zhou, and Hong Fan. Reduced miR-29a-3p expression is linked to the cell proliferation and cell migration in gastric cancer. *World Journal of Surgical Oncology*, 13(1):101, 2015.

[243] Anna Torres, Kamil Torres, Anna Pesci, Marcello Ceccaroni, Tomasz Paszkowski, Paola Cassandrini, Giuseppe Zamboni, and Ryszard Maciejewski. Deregulation of miR-100, miR-99a and miR-199b in tissues and plasma coexists with increased expression of mTOR kinase in endometrioid endometrial carcinoma. *BMC Cancer*, 12(1):369, 2012.

[244] Mor Hanan, Hermona Soreq, and Sebastian Kadener. CircRNAs in the brain. *RNA Biology*, 14(8):1028–1034, 2017.

[245] William R Jeck and Norman E Sharpless. Detecting and characterizing circular RNAs. *Nature Biotechnology*, 32(5):453–461, 2014.

[246] Dongming Liang, Deirdre C Tatomer, Zheng Luo, Huang Wu, Li Yang, Ling-Ling Chen, Sara Cherry, and Jeremy E Wilusz. The output of protein-coding genes shifts to circular RNAs when the pre-mRNA processing machinery is limiting. *Molecular Cell*, 68(5):940–954, 2017.

[247] Thomas B Hansen, Trine I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, and Jørgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441):384–388, 2013.

[248] Julia Salzman, Raymond E Chen, Mari N Olsen, Peter L Wang, and Patrick O Brown. Cell-type specific features of circular RNA expression. *PLoS Genetics*, 9(9):e1003777, 2013.

[249] Agnieszka Rybak-Wolf, Christin Stottmeister, Petar Glažar, Marvin Jens, Natalia Pino, Sebastian Giusti, Mor Hanan, Mikaela Behm, Osnat Bartok, Reut Ashwal-Fluss, et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Molecular Cell*, 58(5):870–885, 2015.

[250] Morten T Venø, Thomas B Hansen, Susanne T Venø, Bettina H Clausen, Manuela Grebing, Bente Finsen, Ida E Holm, and Jørgen Kjems. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. *Genome Biology*, 16(1):245, 2015.

[251] Monika Piwecka, Petar Glažar, Luis R Hernandez-Miranda, Sebastian Memczak, Susanne A Wolf, Agnieszka Rybak-Wolf, Andrei Filipchyk, Filippos Klironomos, Cledi Alicia Cerda Jara, Pascal Fenske, et al. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science*, 357(6357):eaam8526, 2017.

[252] Simon J Conn, Katherine A Pillman, John Toubia, Vanessa M Conn, Marika Salmanidis, Caroline A Phillips, Suraya Roslan, Andreas W Schreiber, Philip A Gregory, and Gregory J Goodall. The RNA binding protein quaking regulates formation of circRNAs. *Cell*, 160(6):1125–1134, 2015.

[253] Nagarjuna Reddy Pamudurti, Osnat Bartok, Marvin Jens, Reut Ashwal-Fluss, Christin Stottmeister, Larissa Ruhe, Mor Hanan, Emanuel Wyler, Daniel Perez-Hernandez, Evelyn Ramberger, et al. Translation of circRNAs. *Molecular Cell*, 66(1):9–21, 2017.

[254] Ivano Legnini, Gaia Di Timoteo, Francesca Rossi, Mariangela Morlando, Francesca Briganti, Olga Sthandier, Alessandro Fatica, Tiziana Santini, Adrian Andronache, Mark Wade, et al. Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Molecular Cell*, 66(1):22–37, 2017.

[255] Blanche Capel, Amanda Swain, Silvia Nicolis, Adam Hacker, Michael Walter, Peter Koopman, Peter Goodfellow, and Robin Lovell-Badge. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell*, 73(5):1019–1030, 1993.

[256] Fang Li, Liyuan Zhang, Wei Li, Jieqiong Deng, Jian Zheng, Mingxing An, Jiachun Lu, and Yifeng Zhou. Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/β-catenin pathway. *Oncotarget*, 6(8):6001, 2015.

[257] Yang Zhang, Xiao-Ou Zhang, Tian Chen, Jian-Feng Xiang, Qing-Fei Yin, Yu-Hang Xing, Shanshan Zhu, Li Yang, and Ling-Ling Chen. Circular intronic long noncoding RNAs. *Molecular Cell*, 51(6):792–806, 2013.

[258] S Lindquist and EA Craig. The heat-shock proteins. *Annual Review of Genetics*, 22(1):631–677, 1988.

[259] Luke Whitesell and Susan L Lindquist. HSP90 and the chaperoning of cancer. *Nature Reviews Cancer*, 5(10):761, 2004.

[260] Mikko Taipale, Daniel F Jarosz, and Susan Lindquist. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature Reviews Molecular Cell Biology*, 11(7):515–528, 2010.

[261] Suzanne L Rutherford and Susan Lindquist. Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709):336–342, 1998.

[262] Vincent Sollars, Xiangyi Lu, Li Xiao, Xiaoyan Wang, Mark D Garfinkel, and Douglas M Ruden. Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nature Genetics*, 33(1):70–74, 2003.

[263] Leah E Cowen and Susan Lindquist. Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. *Science*, 309(5744):2185–2189, 2005.

[264] Christine Queitsch, Todd A Sangster, and Susan Lindquist. Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889):618–624, 2002.

[265] Michael Johnston, Marie-Claude Geoffroy, Andrew Sobala, Ron Hay, and Gyorgy Hutvagner. HSP90 protein stabilizes unloaded argonaute complexes and microscopic P-bodies in human cells. *Molecular Biology of the Cell*, 21(9):1462–1469, 2010.

[266] Justin M Pare, Nasser Tahbaz, Joaquín López-Orozco, Paul LaPointe, Paul Lasko, and Tom C Hobman. Hsp90 regulates the function of argonaute 2 and its recruitment to stress granules and P-bodies. *Molecular Biology of the Cell*, 20(14):3273–3284, 2009.

[267] Tomohiro Miyoshi, Akiko Takeuchi, Haruhiko Siomi, and Mikiko C Siomi. A direct role for Hsp90 in pre-RISC formation in Drosophila. *Nature Structural & Molecular Biology*, 17(8):1024–1026, 2010.

[268] Asha A Nair, Nifang Niu, Xiaojia Tang, Kevin J Thompson, Liewei Wang, Jean-Pierre Kocher, Subbaya Subramanian, and Krishna R Kalari. Circular RNAs and their associations with breast cancer subtypes. *Oncotarget*, 7(49):80967, 2016.

[269] Hani Choudhry, Ashwag Albukhari, Matteo Morotti, Syed Hider, Daniela Moralli, James Smythies, Johannes Schödel, Catherine M Green, Carme Camps, Francesca Buffa, et al. Tumor hypoxia induces nuclear paraspeckle formation through HIF-2α dependent transcriptional activation of NEAT1 leading to cancer cell survival. *Oncogene*, 2015.

[270] Stephen B Edge and Carolyn C Compton. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology*, 17(6):1471–1474, 2010.

[271] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

[272] F Krueger. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 2015.

[273] Simon Andrews et al. FastQC: a quality control tool for high throughput sequence data. 2010.

[274] Yuan Gao, Jinfeng Wang, and Fangqing Zhao. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biology*, 16(1):4, 2015.

[275] Kai Wang, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, Piotr Mieczkowski, Sara A Grimm, Charles M Perou, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178–e178, 2010.

[276] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[277] A Bonizzato, E Gaffo, G Te Kronnie, and S Bortoluzzi. CircRNAs in hematopoiesis and hematological malignancies. *Blood Cancer Journal*, 6(10):e483, 2016.

[278] Mohammad Ali Faghihi, Farzaneh Modarresi, Ahmad M Khalil, Douglas E Wood, Barbara G Sahagan, Todd E Morgan, Caleb E Finch, Georges St Laurent III, Paul J Kenny, and Claes Wahlestedt. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nature Medicine*, 14(7):723–730, 2008.

[279] Xiaowei Wang. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, 32(9):1316–1322, 2016.

[280] Eva Szegezdi, Susan E Logue, Adrienne M Gorman, and Afshin Samali. Mediators of endoplasmic reticulum stress-induced apoptosis. *EMBO Reports*, 7(9):880–885, 2006.

[281] Yu-Ying Hsieh, Po-Hsiang Hung, and Jun-Yi Leu. Hsp90 regulates nongenetic variation in response to environmental stress. *Molecular Cell*, 50(1):82–92, 2013.

[282] Thomas B Hansen, Morten T Venø, Christian K Damgaard, and Jørgen Kjems. Comparison of circular RNA prediction tools. *Nucleic Acids Research*, 44(6):e58–e58, 2015.

[283] John J Rossi. A novel nuclear miRNA mediated modulation of a non-coding antisense RNA and its cognate sense coding mRNA. *The EMBO Journal*, 30(21):4340–4341, 2011.

[284] Connie Wu, Jessica So, Brandi N Davis-Dusenbery, Hank H Qi, Donald B Bloch, Yang Shi, Giorgio Lagna, and Akiko Hata. Hypoxia potentiates microRNA-mediated gene silencing through posttranslational modification of Argonaute2. *Molecular and Cellular Biology*, 31(23):4760–4774, 2011.

[285] Shogo Tokumaru, Motoshi Suzuki, Hideki Yamada, Masato Nagino, and Takashi Takahashi. let-7 regulates Dicer expression and constitutes a negative feedback loop. *Carcinogenesis*, 29(11):2073–2077, 2008.

[286] Reut Ashwal-Fluss, Markus Meyer, Nagarjuna Reddy Pamudurti, Andranik Ivanov, Osnat Bartok, Mor Hanan, Naveh Evantal, Sebastian Memczak, Nikolaus Rajewsky, and Sebastian Kadener. circRNA biogenesis competes with pre-mRNA splicing. *Molecular Cell*, 56(1):55–66, 2014.

[287] Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(6):S9, 2007.

[288] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[289] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[290] Andrew C Oates, Luis G Morelli, and Saúl Ares. Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. *Development*, 139(4):625–639, 2012.

[291] Mark Chaplain, Mariya Ptashnyk, and Marc Sturrock. Hopf bifurcation in a gene regulatory network model: Molecular movement causes oscillations. *Mathematical Models and Methods in Applied Sciences*, 25(06):1179–1215, 2015.

[292] Stefano Volinia, Marco Galasso, Stefan Costinean, Luca Tagliavini, Giacomo Gamberoni, Alessandra Drusco, Jlenia Marchesini, Nicoletta Mascellani, Maria Elena Sana, Ramzey Abu Jarour, et al. Reprogramming of miRNA networks in cancer and leukemia. *Genome Research*, 20(5):589–599, 2010.

[293] Bríd M Ryan, Ana I Robles, and Curtis C Harris. Genetic variation in microRNA networks: the implications for cancer research. *Nature Reviews Cancer*, 10(6):389–402, 2010.

[294] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97, 2013.

[295] John Tsang, Jun Zhu, and Alexander van Oudenaarden. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular Cell*, 26(5):753–767, 2007.

[296] Xin Lai, Olaf Wolkenhauer, and Julio Vera. Understanding microRNA-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Research*, 44(13):6019–6035, 2016.

[297] Baltazar D Aguda, Yangjin Kim, Melissa G Piper-Hunter, Avner Friedman, and Clay B Marsh. MicroRNA regulation of a cancer network: consequences of the feedback loops involving miR-17-92, E2F, and Myc. *Proceedings of the National Academy of Sciences*, 105(50):19678–19683, 2008.

[298] Yichen Li, Yumin Li, Hui Zhang, and Yong Chen. MicroRNA-mediated positive feedback loop and optimized bistable switch in a cancer network involving miR-17-92. *PLoS One*, 6(10):e26302, 2011.

[299] Mingyang Lu, Mohit Kumar Jolly, Herbert Levine, José N Onuchic, and Eshel Ben-Jacob. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proceedings of the National Academy of Sciences*, 110(45):18144–18149, 2013.

[300] Richard Moore, Hsu Kiang Ooi, Taek Kang, Leonidas Bleris, and Lan Ma. MiR-192-mediated positive feedback loop controls the robustness of stress-induced p53 oscillations in breast cancer cells. *PLoS Computational Biology*, 11(12):e1004653, 2015.

[301] Daniel W Thomson and Marcel E Dinger. Endogenous microRNA sponges: evidence and controversy. *Nature Reviews Genetics*, 17(5):272–283, 2016.

[302] Maria D Paraskevopoulou, Georgios Georgakilas, Nikos Kostoulas, Martin Reczko, Manolis Maragkakis, Theodore M Dalamagas, and Artemis G Hatzigeorgiou. DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Research*, 41(D1):D239–D245, 2012.

[303] Ashwini Jeggari, Debora S Marks, and Erik Larsson. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics*, 28(15):2062–2063, 2012.

[304] Margaret S Ebert and Phillip A Sharp. MicroRNA sponges: progress and possibilities. *RNA*, 16(11):2043–2050, 2010.

[305] Jeff Hasty, David McMillen, Farren Isaacs, and James J Collins. Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics*, 2(4):268–279, 2001.

[306] Leon Glass and Stuart A Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, 1973.

[307] Leon Glass. Classification of biological networks by their qualitative dynamics. *Journal of Theoretical Biology*, 54(1):85–107, 1975.

[308] Ting Chen, Hongyu L He, and George M Church. Modeling gene expression with differential equations. In *Biocomputing'99*, pages 29–40. World Scientific, 1999.

[309] Tianhai Tian, Kevin Burrage, Pamela M Burrage, and Margherita Carletti. Stochastic delay differential equations for genetic regulatory networks. *Journal of Computational and Applied Mathematics*, 205(2):696–707, 2007.

[310] Harley H McAdams, Lucy Shapiro, et al. Circuit simulation of genetic networks. *Science*, 269(5224):650–656, 1995.

[311] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[312] Moises Santillán. On the use of the Hill functions in mathematical models of gene regulatory networks. *Mathematical Modelling of Natural Phenomena*, 3(2):85–97, 2008.

[313] Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[314] Albert Goldbeter. A model for circadian oscillations in the Drosophila period protein (PER). *Proceedings of the Royal Society of London B: Biological Sciences*, 261(1362):319–324, 1995.

[315] Chunguang Li and Guanrong Chen. Synchronization in general complex dynamical networks with coupling delays. *Physica A: Statistical Mechanics and its Applications*, 343:263–278, 2004.

[316] Luonan Chen and Kazuyuki Aihara. Stability of genetic regulatory networks with time delay. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(5):602–608, 2002.

[317] Leon Glass. Synchronization and rhythmic processes in physiology. *Nature*, 410(6825):277–284, 2001.

[318] Arthur T Winfree. Biological rhythms and the behavior of populations of coupled oscillators. *Journal of Theoretical Biology*, 16(1):15–42, 1967.

[319] Renato E Mirollo and Steven H Strogatz. Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, 50(6):1645–1662, 1990.

[320] Matthias B Wahl, Chuxia Deng, Mark Lewandoski, and Olivier Pourquié. FGF signaling acts upstream of the NOTCH and WNT signaling pathways to control segmentation clock oscillations in mouse somitogenesis. *Development*, 134(22):4033–4041, 2007.

[321] Katrin Serth, Karin Schuster-Gossler, Ralf Cordes, and Achim Gossler. Transcriptional oscillation of lunatic fringe is essential for somitogenesis. *Genes & Development*, 17(7):912–925, 2003.

[322] Jacqueline Kim Dale, Pascale Malapert, Jérome Chal, Gonçalo Vilhais-Neto, Miguel Maroto, Teri Johnson, Sachintha Jayasinghe, Paul Trainor, Bernhard Herrmann, and Olivier Pourquié. Oscillations of the snail genes in the presomitic

mesoderm coordinate segmental patterning and morphogenesis in vertebrate somitogenesis. *Developmental Cell*, 10(3):355–366, 2006.

[323] Mary-Lee Dequéant, Earl Glynn, Karin Gaudenz, Matthias Wahl, Jie Chen, Arcady Mushegian, and Olivier Pourquié. A complex oscillating network of signaling genes underlies the mouse segmentation clock. *science*, 314(5805):1595–1598, 2006.

[324] Olivier Pourquié. The segmentation clock: converting embryonic time into spatial pattern. *Science*, 301(5631):328–330, 2003.

[325] John Cooke and Erik Christopher Zeeman. A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *Journal of theoretical biology*, 58(2):455–476, 1976.

[326] Yun-Jin Jiang, Birgit L Aerne, Lucy Smithers, Catherine Haddon, David Ish-Horowicz, and Julian Lewis. Notch signalling and the synchronization of the somite segmentation clock. *Nature*, 408(6811):475, 2000.

[327] Ruth E Baker, S Schnell, and PK Maini. A clock and wavefront mechanism for somite formation. *Developmental Biology*, 293(1):116–126, 2006.

[328] Leah Herrgen, Saúl Ares, Luis G Morelli, Christian Schröter, Frank Jülicher, and Andrew C Oates. Intercellular coupling regulates the period of the segmentation clock. *Current Biology*, 20(14):1244–1253, 2010.

[329] David K Welsh, Diomedes E Logothetis, Markus Meister, and Steven M Reppert. Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms. *Neuron*, 14(4):697–706, 1995.

[330] Albert Goldbeter. Computational approaches to cellular rhythms. *Nature*, 420(6912):238–245, 2002.

[331] Steven H Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1-4):1–20, 2000.

[332] Chen Liu, David R Weaver, Steven H Strogatz, and Steven M Reppert. Cellular construction of a circadian clock: period determination in the suprachiasmatic nuclei. *Cell*, 91(6):855–860, 1997.

[333] Béla Novák and John J Tyson. Design principles of biochemical oscillators. *Nature Reviews Molecular Cell Biology*, 9(12):981–991, 2008.

[334] Lan Ma, John Wagner, John Jeremy Rice, Wenwei Hu, Arnold J Levine, and Gustavo A Stolovitzky. A plausible model for the digital response of p53 to DNA damage. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14266–14271, 2005.

[335] Nicholas AM Monk. Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptional time delays. *Current Biology*, 13(16):1409–1413, 2003.

[336] DE Nelson, AEC Ihekwaba, M Elliott, JR Johnson, CA Gibney, BE Foreman, G Nelson, V See, CA Horton, DG Spiller, et al. Oscillations in NF-κB signaling control the dynamics of gene expression. *Science*, 306(5696):704–708, 2004.

[337] Jean-Louis Martiel and Albert Goldbeter. A model based on receptor desensitization for cyclic AMP signaling in Dictyostelium cells. *Biophysical Journal*, 52(5):807–828, 1987.

[338] Albert Goldbeter. Mechanism for oscillatory synthesis of cyclic AMP in Dictyostelium discoideum. *Nature*, 253(5492):540–542, 1975.

[339] Peter Ruoff, Merete Vinsjevik, Christian Monnerjahn, and Ludger Rensing. The Goodwin oscillator: on the importance of degradation reactions in the circadian clock. *Journal of Biological Rhythms*, 14(6):469–479, 1999.

[340] Andrew Dhawan, Abdullah Hamadeh, and Brian Ingalls. Designing synchronization protocols in networks of coupled nodes under uncertainty. In *American Control Conference (ACC), 2012*, pages 4945–4950. IEEE, 2012.

[341] Michael G Rosenblum, Arkady S Pikovsky, and Jürgen Kurths. Phase synchronization of chaotic oscillators. *Physical Review Letters*, 76(11):1804, 1996.

[342] Yoshiki Kuramoto and Ikuko Nishikawa. Statistical macrodynamics of large dynamical systems. Case of a phase transition in oscillator communities. *Journal of Statistical Physics*, 49(3-4):569–605, 1987.

[343] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3):353–358, 2011.

[344] Marvin Jens and Nikolaus Rajewsky. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nature Reviews Genetics*, 16(2):113–126, 2015.

[345] Yvonne Tay, John Rinn, and Pier Paolo Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–352, 2014.

[346] Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301):1033–1038, 2010.

[347] Amanda N Kallen, Xiao-Bo Zhou, Jie Xu, Chong Qiao, Jing Ma, Lei Yan, Lingeng Lu, Chaochun Liu, Jae-Sung Yi, Haifeng Zhang, et al. The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Molecular Cell*, 52(1):101–112, 2013.

[348] Fritz Horn and Roy Jackson. General mass action kinetics. *Archive for Rational Mechanics and Analysis*, 47(2):81–116, 1972.

[349] Brian Ingalls, Maya Mincheva, and Marc R Roussel. Parametric sensitivity analysis of oscillatory delay systems with an application to gene regulation. *Bulletin of Mathematical Biology*, pages 1–25, 2017.

[350] Edward John Routh. *A treatise on the stability of a given state of motion: particularly steady motion.* Macmillan and Company, 1877.

[351] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337, 2011.

[352] Michael Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

[353] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[354] Long Cai, Chiraj K Dalal, and Michael B Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(7212):485, 2008.

[355] Irfan A Qureshi and Mark F Mehler. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nature Reviews Neuroscience*, 13(8):528, 2012.

[356] Shuohao Liao, Tomáš Vejchodskỳ, and Radek Erban. Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. *Journal of The Royal Society Interface*, 12(108):20150233, 2015.

[357] Tjeerd olde Scheper, Don Klinkenberg, Cyriel Pennartz, and Jaap Van Pelt. A mathematical model for the intracellular circadian rhythm generator. *Journal of Neuroscience*, 19(1):40–47, 1999.

[358] Dmitri Bratsun, Dmitri Volfson, Lev S Tsimring, and Jeff Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14593–14598, 2005.

[359] Tobias Galla. Intrinsic fluctuations in stochastic delay systems: Theoretical description and application to a simple model of gene regulation. *Physical Review E*, 80(2):021909, 2009.

[360] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[361] David F Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of Chemical Physics*, 127(21):214107, 2007.

[362] Yehoshua Enuka, Mattia Lauriola, Morris E Feldman, Aldema Sas-Chen, Igor Ulitsky, and Yosef Yarden. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. *Nucleic Acids Research*, 44(3):1370–1383, 2016.

[363] Hannah Gruner, Mariela Cortés-López, Daphne A Cooper, Matthew Bauer, and Pedro Miura. CircRNA accumulation in the aging mouse brain. *Scientific Reports*, 6:38907, 2016.

[364] Michael B Clark, Rebecca L Johnston, Mario Inostroza-Ponta, Archa H Fox, Ellen Fortini, Pablo Moscato, Marcel E Dinger, and John S Mattick. Genome-wide analysis of long noncoding RNA stability. *Genome Research*, 22(5):885–898, 2012.

[365] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335, 2000.