NOVEL COMPUTATIONAL APPROACHES TO RESEARCH LONGITUDINAL MICRORNA-MRNA EXPRESSION DATASETS

A thesis submitted by

Krutik Patel

For the Degree of Doctor of Philosophy

Newcastle University Biosciences Institute

Newcastle University

September 2021

Abstract

microRNAs (miRNAs) regulate many biological processes and are used as biomarkers for the classification of diseases, conditions and developmental stages. miRNAs function by targeting and negatively regulating specific mRNAs. One limitation of utilising miRNAs in experimental work is the complex and often redundant behaviour of miRNA-mRNA interactions; as a single miRNA can regulate many mRNAs and one mRNA can be regulated by multiple miRNAs. This complexity stifles the potential of miRNAs. However, miRNA-mRNA expression datasets are becoming generated more frequently and they can help to garner greater understanding of how miRNAs regulate biological systems. Furthermore, researchers are generating longitudinal datasets as these can elude to greater understanding of how biological conditions change over time. Thus there is a rise of longitudinal miRNA-mRNA expression datasets. However, extracting useful information from increasingly sophisticated datasets is a challenge in biological research. Exploration of such datasets using computational techniques, such as big data bioinformatics, kinetic modelling and machine learning could help in identifying interesting miRNA-mRNA interactions. During this PhD I asked if these methodologies can be used to gain insights from a range of longitudinal miRNA-mRNA expression datasets. Hence, I developed an *R/Bioconductor* tool called *TimiRGeN* to integrate, analyse and generate small networks from longitudinal miRNA-mRNA datasets. Datasets from kidney fibrosis, chondrogenesis dataset, breast cancer and Huntington's disease (HD) were analysed with *TimiRGeN*. Results from the chondrogenesis dataset analysis were taken forward to generate a multi-miRNA kinetic model. With help from my collaborators this model was validated and predictions were made. Using the HD dataset, machine learning (ML) techniques trained models to detect if samples have disease or wild type conditions. Overall, I have developed and used multiple computational techniques to increase knowledge gained from longitudinal miRNA-mRNA datasets, and I believe the results show these techniques can contribute to miRNA research.

Acknowledgments

I would like to thank my supervisors Daryl Shanley, David Young and Carole Proctor who have guided me through the last four years. This supervisory team aided me with technical, financial and emotional support throughout my PhD. I especially want to give credit to Daryl for shepherding me into becoming a more independent researcher.

I am grateful to Ian Clark, David Young and the Dunhill medical group for establishing the funding for this PhD project.

I would also like to thank my experimental collaborator Matt Barter who generated excellent data to be used in my PhD.

A special thank you to those in my group who helped me with technical and academic help, and they to were just great company during this PhD: Ciaran Welsh, Alvaro Martinez Guimera, Sharmilla Chandrasegaran, James Wordsworth, Louise Pease, Jamie Soul, Karthyn Garner, Peter Clark and Gina Abdelaal. Also a thanks to project students Bethany Harley and Colleen Sheridan.

Finally, I would like to thank the *Bioconductor* and stack exchange community for helping me with technical issues that I encountered.

Declaration

I declare the work presented within this thesis is based my own research and has not been submitted as part of another degree at any institution. Information derived from published work has been cited in the bibliography.

List of Figures

List of Tables

List of Abbreviations

c mode = combined mode

cAMP-PKA =Cyclic AMP dependent protein kinase A

ceRNA = competing endogenous RNA

CI = Confidence Intervals

Ch = Chapter

CSF = Cerebrospinal Fluid

D = Days

DE = differentially expressed

eIF = Eukaryotic Initiation Factor

ECM = extracellular matrix

FA = Folic acid

GAG = glycosaminoglycan

GEO = gene expression omnibus

GRN = Gene Regulatory Network

HD = Huntington's disease

hp = hairpin

IP = Internet Protocol

IRP = Inflammatory Response Pathway

JHD = juvenile onset Huntington's disease

kDa = killodalton

KO = knock out

LRT = Likelihood ratio test

M = Month

MAE = MultiAssayExperiment

miR-199a/b-5p = miR-199a-5p or miR-199b-5p

miRNAs = microRNAs

mHTT = mutant Huntington's gene/ protein

ML = Machine Learning

MSCs = mesenchymal stem cells

nt = nucleotides

ODEs = ordinary differential expressions

ORA = overrepresentation analysis

poly-Q = poly-Glutamine

s mode = separated mode

SDEGs = significantly differentially expressed genes

TDMD = target-directed miRNA degradation

UTR = untranslated region

UUO = Unilateral Ureter Obstruction

wtHTT = wild type Huntington's gene/ protein

Nomenclature

Human: PROTEIN, *MRNA*, *GENE*, *hsa-mirna*

Mouse: PROTEIN, *Mrna*, *Gene*, *mmu-mirna*

Generic references to proteins, mRNAs, genes and miRNAs will be written with human nomenclature.

Generic references to miRNAs will not specific *hsa/mmu* prefixes.

During modelling/ GRN design naming conventions are dropped as I will be referring to model species/ components.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 microRNAs

1.1.1 microRNA biogenesis

microRNAs (miRNAs) are small non-coding single stranded pieces of RNA, roughly 16-22 nucleotides (nt) long. They post transcriptionally negatively modulate gene expression of specific mRNAs [1, 2, 3]. Binding of a mature miRNA to its target mRNA occurs between a 7-8 nt sized region found in the 5' end of the miRNA, otherwise known as the seed sequence, and complementary binding sites, most often found on the 3' UTR of the target mRNA (Figure 1.1) [4, 5, 6]. A mature miRNA must undergo a multi-step and tightly regulated biogenesis process that begins with transcription by RNA polymerase II to form a hairpin shaped double stranded piece of RNA known as a pri-miRNA [7, 8]. Nuclear RNase III DROSHA and its co-factor DGCR8 process the pri-miRNA in the nucleus to form a 70-100 nt long pre-miRNA [9]. Specific sequences of the pre-mRNA is recognised by nuclear export protein EXPORTIN5. This protein will then export the pre-miRNA into the cytoplasm via a RAN-GTP dependent manner [9, 10, 11, 12]. EXPORTIN5 will be attached to a GTP molecule, and upon recognising its cargo (a pre-miRNA molecule), hydrolysis will turn the GTP into a GDP, and this leads to a conformational change in EXPORTIN5. EXPORTIN5 will be recycled for further use [13].

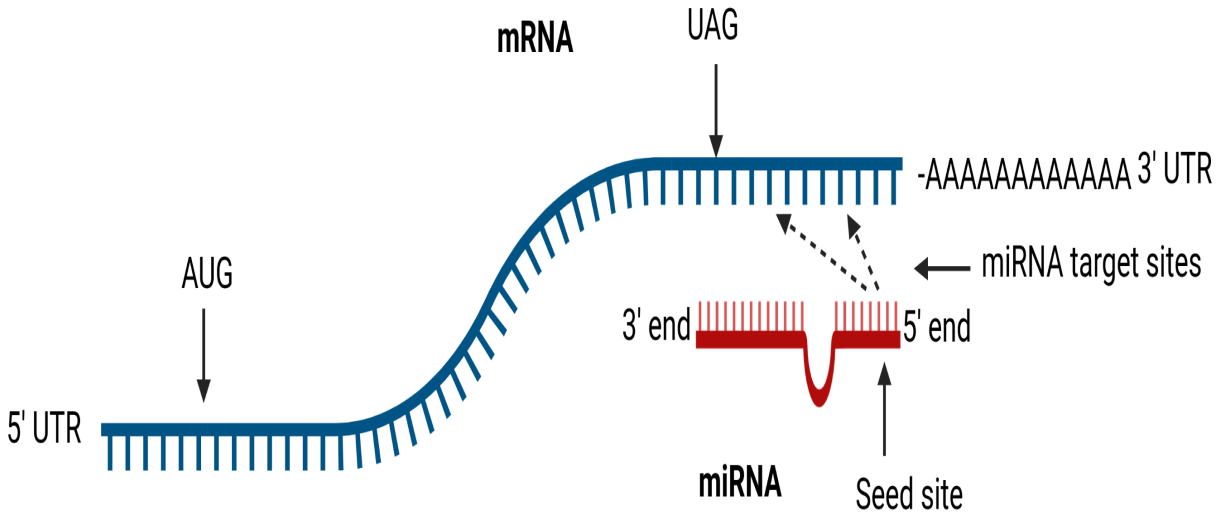


Figure 1.1: **miRNA-mRNA interaction illustration**. The figure shows a basic schematic of how a miRNA-mRNA interaction works, and also provides some detail on how the seed site region of a mature miRNA can target multiple regions of the 3' UTR region of a mRNA. The AUG sequence is the start codon and the UAG sequence is a stop codon. A poly-A tail is found at the end of the 3' UTR, and the 5' end of the miRNA binds to the 3' end of the mRNA, before the poly-A tail. The miRNA may also have multiple target sequences on the 3' UTR of the mRNA.

Cytoplasmic RNase DICER processes the pre-miRNA into a shorter miRNA duplex [14]. DICER associated protein TRBP recruits a protein complex called RISC to begin strand selection [15]. One strand of the duplex becomes the mature guide miRNA, and is incorporated into the RISC protein complex, forming a miRISC complex (further explained in subsection 1.1.3) [15]. The latter strand, known as the carrier strand is frequently degraded [16]. The guide miRNA leads the miRISC complex to target mRNAs for degradation or translational inhibition. The miRISC complex serves as a miRNA induced gene silencing unit, and is made up of many proteins, but the core proteins are an Argonaute2 (AGO2) protein and a member of the GW182 family of proteins (Figure 1.2A - 1.2C) [17].

miRtrons

The pre-processing steps described above is not uniform because non-canonical biogen-

esis of miRNAs has also been reported [18]. These have been referred to as miRtrons. miRtrons are transcribed from introns and are then spliced as pre-miRNAs (Figure 1.2D). They do not possess binding motifs for DROSHA recognition; and instead bypass DROSHA to be exported out of the nucleus for processing by cytoplasmic DICER [19, 20]. Pre-miRNAs have also been reported to be derived from snoRNAs and tRNAs [21, 22].

1.1.2 miRISC complex

AGO2

The RISC complex and a mature miRNA make up the miRISC complex which will use the miRNA to locate target mRNAs for silencing. The core of the RISC complex comprises of an AGO2 unit and a GW182 unit. Mammals have four AGO proteins, and of which only AGO2 has a catalytic domain [17, 23]. AGO proteins are large and have multiple domains such as: the N domain which can recognise, recruit and unwind duplexes of RNAs, the PAZ domain which is where the 5' end of the miRNA is pocketed, the MID domain which is where the 3' end of a mature miRNA is bound, and there is a PIWI domain. In the case for AGO2, the PIWI domain is the major source of endonuclease activity in the miRISC complex [24]. A further function of the AGO2 subunit is to protect the ends of a mature miRNA from degradation, and this contributes to why some miRNAs have been found to have very long half lives [25, 26]. Often due to complementary mis-matches the miRNA-mRNA interactions can be misaligned, making mRNA degradation by AGO2 induced endonuclease activity imprecise. More commonly, miRNA induced silencing is a result of deadenylation, decapping and 5'-3' decay, which is orchestrated by the AGO2 associated protein from the GW182 family.

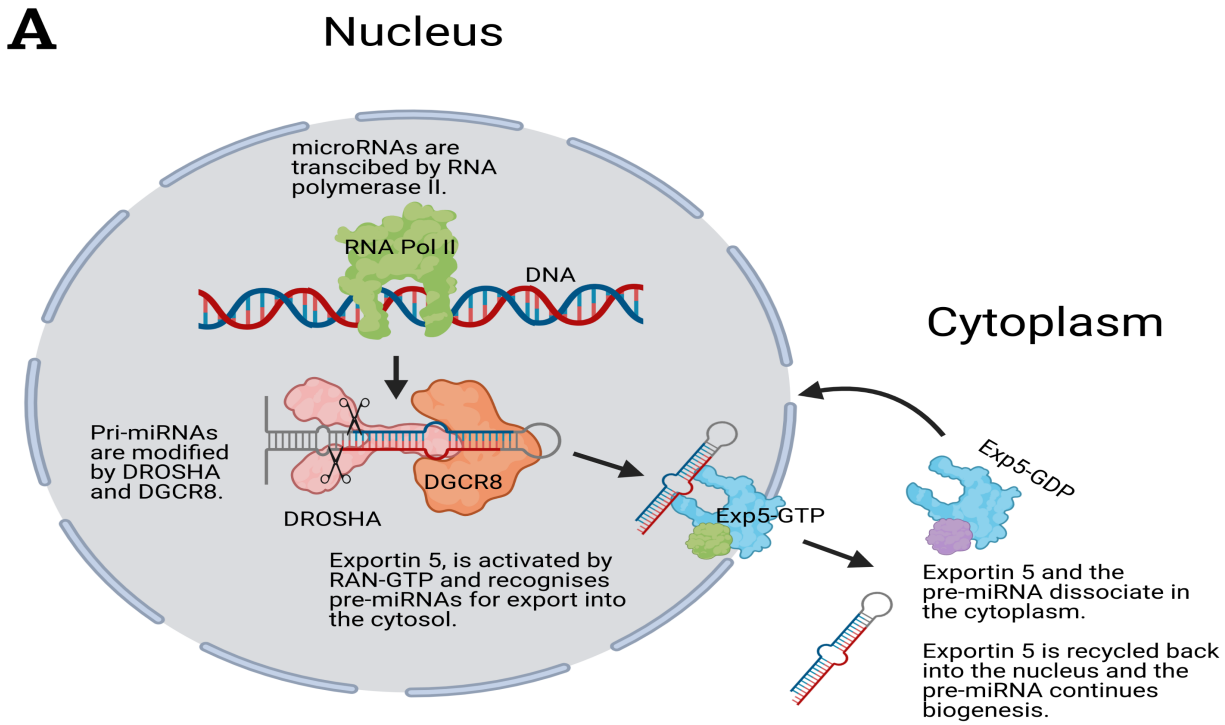
GW182 protein family

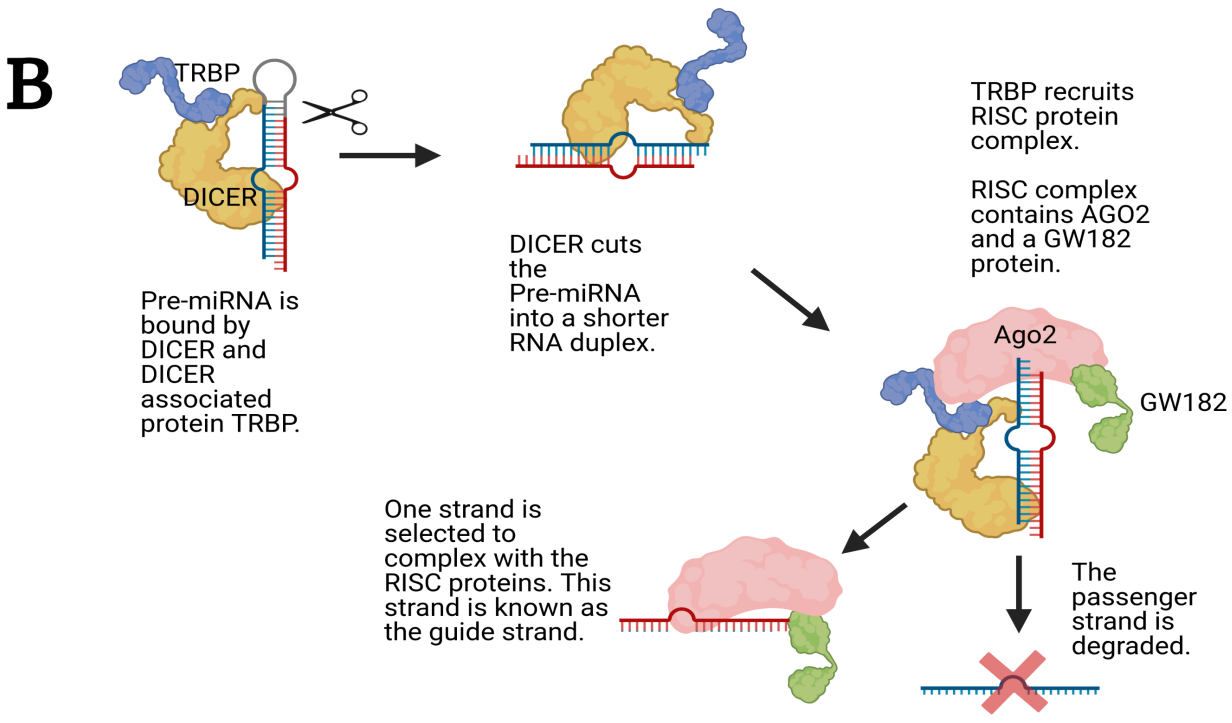
GW182 protein family includes TNRC6A, TNRC6B and TNRC6C. Within a RISC complex, the N-terminus of a GW182 protein will directly be bound to the PIWI domain of the AGO2 sub-unit [27, 28]. The AGO2 unit is attracted to the multiple Gly-Trp (GW) repeats found in the N-terminus [29]. GW182 proteins have a flexible region and this is tailed by a C-terminus which contains an RRM region which has the ability to bind with PABPC to induce

transcriptional regulation [29]. PABPC is a highly conserved eukaryotic protein which is essential for translation initiation and mRNA stability. PABPC is bound to the poly(A) tail on the 3' end of mRNAs, protecting that end and circularising the mRNA by binding to the eukaryotic CAP/eIF (eukaryotic initiation factor) protein complex [30, 31].

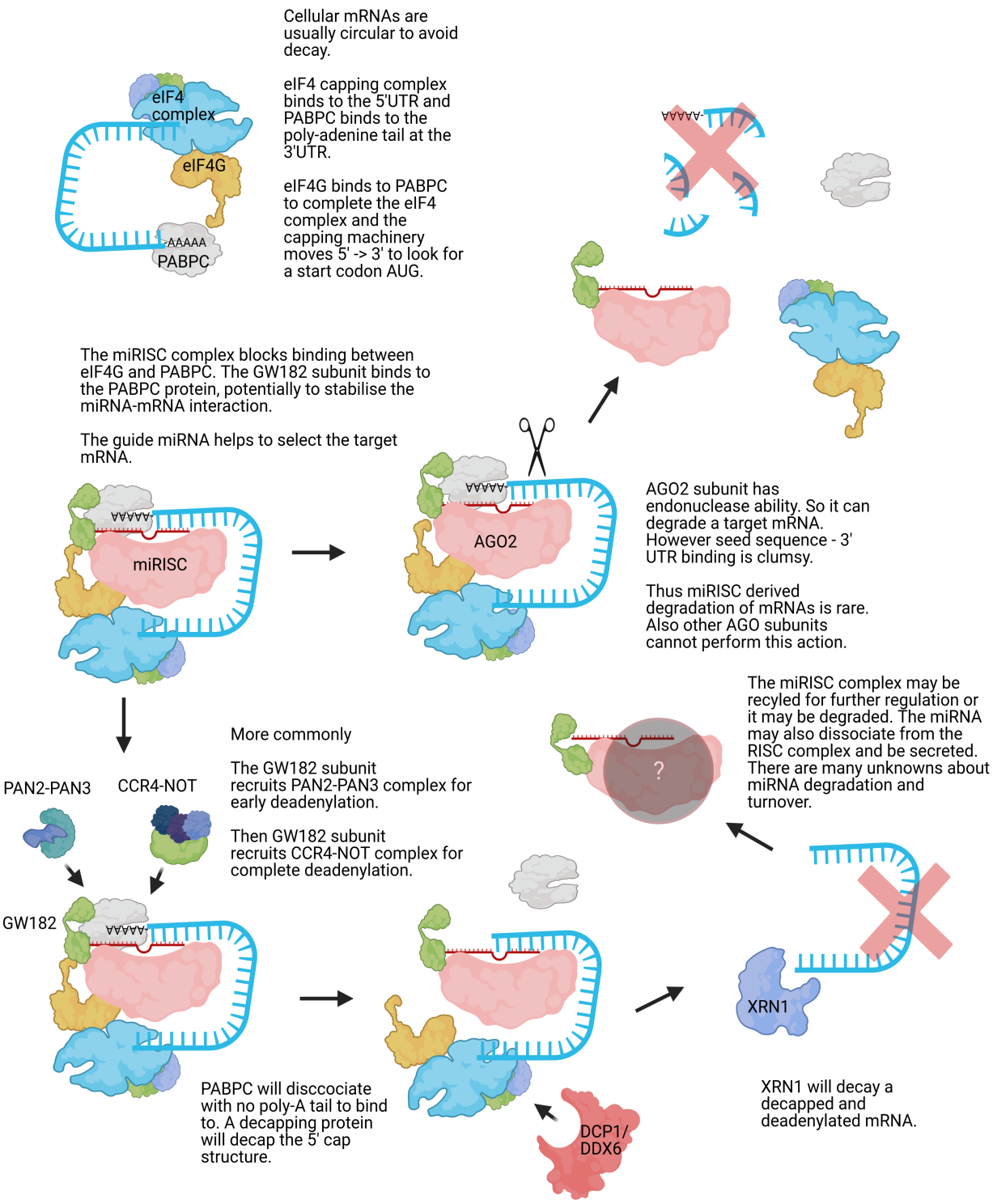
mRNA regulation by the miRISC complex

Commonly, mRNAs are found in a circularised structure, rather than an open structure to avoid decaying enzymes. The circularisation of a mRNA is the result of complexing with PABPC and the eIF4 complex. The resulting protein-mRNA complex will begin the process of translation. However, GW182 subunits compete with eIF4G of the CAP complex, by binding with PABPC [32, 33]. This leads to the mRNA staying in the open structure, blocking translation and increasing the likelihood of mRNA decay [30]. Furthermore, inhibition of PABPC-CAP complex binding reduces the ability of the eIF4 proteins to recognise legitimate mRNAs for translation; reducing the gene expression of mRNAs which are targetted by a miRISC complex [28, 34]. Binding of GW182 proteins and PABPC is useful in co-ordinating miRISC induced transcriptional silencing [35]. GW182 proteins also recruits cytoplasmic deadenylation complex PAN2-PAN3 which will catalyse the early phase of deadenylation by clipping the 3' poly(A) repeat chain to a smaller poly(A) repeat chain [36, 37, 38, 39]. Following this, GW182 recruits another cytoplasmic deadenylase complex, called the CCR4-NOT complex which comprises of NOT1, NOT2, NOT3 and NOT9 [39, 40, 41, 42, 43, 44]. This complex will complete deadenylation, and it has been shown that CCR4-NOT is sufficient to complete decapping without PAN2-PAN3. After this, the targetted mRNA will be decapped by an associated decapping protein e.g. DCP1 or DDX6 [45]. Deadynylated and decapped mRNAs are targets for XRN1 to initiate 5'-3' decay of the targetted mRNA (Figure 2C) [46].


A
Nucleus
microRNAs are transcibed by RNA polymerase II.
RNA Pol II
DNA
Pri-miRNAs are modified by DROSHA and DGCR8.
DGCR8
DROSHA
Exportin 5, is activated by RAN-GTP and recognises pre-miRNAs for export into the cytosol.
Cytoplasm
Exp5-GDP
Exp5-GTP
Exportin 5 and the pre-miRNA dissociate in the cytoplasm.
Exportin 5 is recycled back into the nucleus and the pre-miRNA continues biogenesis.


B
TRBP
DICER
Pre-miRNA is bound by DICER and DICER associated protein TRBP.
DICER cuts the Pre-miRNA into a shorter RNA duplex.
TRBP recruits RISC protein complex.
RISC complex contains AGO2 and a GW182 protein.
Ago2
GW182
One strand is selected to complex with the RISC proteins. This strand is known as the guide strand.
The passenger strand is degraded.

C


Cellular mRNAs are usually circular to avoid decay.
eIF4 complex
eIF4G
AAAAA
PABPC
eIF4 capping complex binds to the 5'UTR and PABPC binds to the poly-adenine tail at the 3'UTR.
eIF4G binds to PABPC to complete the eIF4 complex and the capping machinery moves 5' -> 3' to look for a start codon AUG.
The miRISC complex blocks binding between eIF4G and PABPC. The GW182 subunit binds to the PABPC protein, potentially to stabilise the miRNA-mRNA interaction.
The guide miRNA helps to select the target mRNA.
miRISC
AGO2
AGO2 subunit has endonuclease ability. So it can degrade a target mRNA. However seed sequence - 3' UTR binding is clumsy.
Thus miRISC derived degradation of mRNAs is rare. Also other AGO subunits cannot perform this action.
The miRISC complex may be recyled for further regulation or it may be degraded. The miRNA may also dissociate from the RISC complex and be secreted. There are many unknowns about miRNA degradation and turnover.
?
More commonly
The GW182 subunit recruits PAN2-PAN3 complex for early deadenylation.
Then GW182 subunit recruits CCR4-NOT complex for complete deadenylation.
PAN2-PAN3
CCR4-NOT
GW182
PABPC will disccociate with no poly-A tail to bind to. A decapping protein will decap the 5' cap structure.
DCP1/ DDX6
XRN1
XRN1 will decay a decapped and deadenylated mRNA.

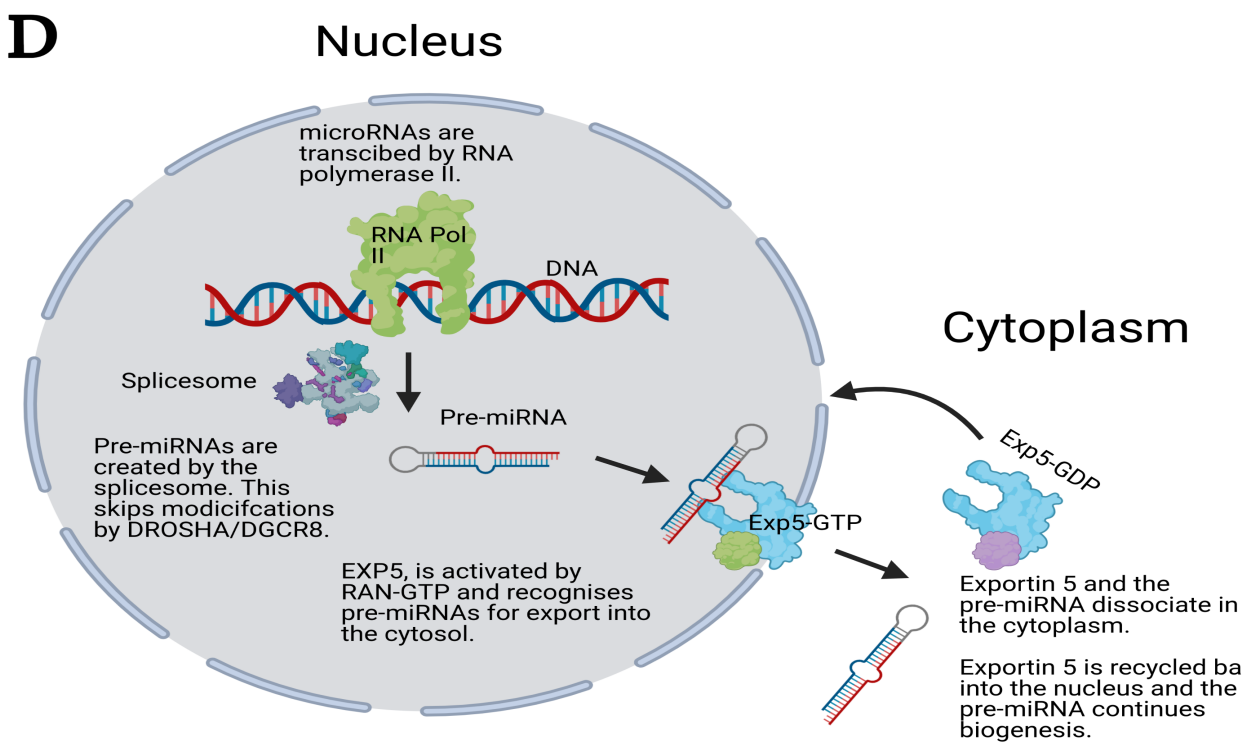


Figure 1.2: **Illustrations of microRNA biogenesis in four stages. A)** details how the miRNA is transcribed by RNA polymerase II and is then modified by a DROSHA-DGCR8 complex to become a pre-miRNA which can be exported to the cytosine by EXPORTIN5. **B)** goes though pre-miRNA maturation into a mature miRNA which complexes with the RISC complex after modulation by DICER-TBPR. **C)** illustrates how miRNA-mRNA targeting and degradation occurs and **D)** shows miRtron biogenesis and here miRNAs skip the pri-miRNA step because they are spliced out of introns as pre-miRNAs. RNA-protein size ratio has not been considered in these images, as proteins would be many times larger than the RNAs.

Overall, AGO2 and GW182 proteins contribute to the miRISC complex's ability to induce mRNA silencing by direct endonuclease activity by AGO2, translation repression by blocking eIF4G-PABPC binding, or by recruitment of deadenylase complexes to deadenylase and then trigger decap the 5' end of the target mRNA, which would lead to subsequent 5'-3' decay. The lattermost method of miRNA silencing method is mostly likely the common approach in higher eukaryotic organisms, but much of this process is still unknown. All in all, miRNA induced mRNA silencing is a very refined method of regulating gene expression [47, 48].

1.1.3 Unknown sections of miRNA biogenesis

There are several unknowns which remain in miRNA biogenesis. One is a complete list of rules for strand selection. So far two rules have been established (nucleotide preference and thermodynamic preference), however not all miRNAs follow these two rules. Nucleotide preference can be broken down into two categories; chemical affinity and phosphorylation. Beginning with the former, it has been shown that AGO2s MID domain preferentially binds to U more than A, G or C nucleotides. Also, A nucleotides are preferentially bound in contrast to G and C. So perhaps G-C content leads to strand selection [49]. Also, it seems phosphorylation of the first nucleotide in a strand can lead to an AGO2 subunit binding to the strand. Blocking phosphorylation at the 5' end of the guide strand leads to random strand selection [50]. The second strand selection rule is that AGO2 subunits seem to preferentially select miRNA strands which are thermodynamically less stable [51]. However, despite these "rules", most miRNAs do not follow them. A comprehensive study by *Medley et al., (2021)* found that in a range of species an average of 24% of miRNAs don't follow a single rule and between 17-24% don't follow both rules [52].

The other biogenesis step which is not fully understood is miRNA decay. The stability of miRNAs vary greatly, some even having half-lives longer than 100 hours [26]. It seems some miRNAs are degraded by target-directed miRNA degradation (TDMD). These are RNAs which complementarily bind to miRNAs to induce miRNA degraded [53]. TDMD RNAs bind to the 3' end of the target miRNA and are hypothesised to alter AGO2 conformation to expose the 3' end of the miRNA, and thus encourages 3'-5' RNA decay of the target miRNA [53, 54, 55]. Though much of this process is still unknown, it may answer how RISC units are recycled and further explain how miRNA turn-over occurs. Alternatively, miRNAs can be excreted into bodily fluids. Some circulating miRNAs are enclosed within exosomes which have the potential to be trafficked to other cells/ tissues [56]. Under specific disease conditions, miRNAs such as *miR-150*, *miR-142-3p*, and *miR-451* are preferentially packaged into circulating exosomes [57]. Furthermore, some miRNAs have been found in to be differentially expressed in disease conditions, for example *miR-21* is

expressed at lower concentrations in serum exosomes of control patients than in serum exosomes of glioblastoma patients. Several miRNA exosome pathways have been described, but little is known about the specifics of which miRNAs are preferentially selected for exosomes, though sequence specificity may be a driving factor in selecting miRNAs to be packaged in circulating exosomes [56]. Another idea is that some circulating miRNAs are a cellular waste products. Most (90%) circulating miRNAs are microvesicle-free and bound to an AGO2 protein [58]. This raises the question, why would cells excrete large, expensive AGO2 proteins? Would it not be more cost-effective for a cell to retain and recycle its AGO2 subunits? Could it be that cells have not developed sophisticated methods to remove unneeded miRNAs from their cells, or to remove miRNAs from AGO2 proteins? TDMD RNAs can degrade some miRNAs, however maybe this is a too time consuming method which cannot efficiently deal with unneeded miRNAs. Thus, cells simply excrete unneeded miRNA-AGO2 complexes. This theory is difficult to test. Regardless of why miRNAs are circulating in biofluids, the fact is they do and can be exploited to learn about the health of individuals.

1.1.4 miRNA-mRNA interactions rules and databases

miRNA target interactions rules

There are many miRNA-mRNA interaction rules which can contribute to the likelihood of a miRNA-mRNA interaction occurring. Some rules are more important than others and contribute to a greater chance of a miRNA-mRNA interaction occurring. Prediction software and databases use many of these rules as features in their algorithms. Such databases provide an essential resource for researchers investigating miRNAs. There are many miRNA-mRNA target databases, and most miRNA-mRNA prediction tools use rule based or machine learning approaches, with a number of biological features to determine the likelihood and strength of predicted miRNA-mRNA interactions. Below I have listed most of the miRNA-mRNA interaction rules and described how each one may affect a miRNA-mRNA interaction [59].

- **Seed site specificity** - The seed sequence of the miRNA (first 2-8 nt from the 5' end). Watson-Crick (A-U, G-C) matches between the seed site and the target mRNA. The closer the nt match, the more likely the miRNA-mRNA will occur and a stronger

interaction is likely to occur if the match is perfect [4]. Many tools use this information in their algorithms, and some allow for a degree of mismatches.

- **Evolutionary conservation** - Many miRNA sequences and mRNA target site sequences have remained evolutionarily conserved for millions of years. It has become an accepted fact that the seed region of the miRNA has far higher conservation across species, than non-seed regions of the miRNA [4, 60]. Some algorithms use conservation between miRNA-mRNA interactions to filter based on cross-species interactions [4]. Generally, the more conserved an interaction is, the more likely the interaction occurs, and some prediction sites reflect this by giving these interactions a better score.

- **Gibbs free energy calculations** - Free energy calculations are used by some software to measure the stability of a predicted miRNA-mRNA interaction.Lower energy releasing interactions are assumed to be more stable, as they have less energy to continue interacting [4, 61]. Some algorithms will have a free energy/ thermodynamic threshold. Overall, free energy calculations can decipher the likelihood of a miRNA-mRNA interaction leading to a stable (favorable) system.

- **mRNA binding site accessibility** - Site accessibility predicts the easy of access a miRNA has to interact with the mRNA binding sites. miRNA-mRNA interactions occur in a two step process: 1) the miRNA binds to a shorter accessible site of the mRNA, 2) the mRNA unfolds to reveal a secondary structure from where the miRNA can access the target sites [62]. Many algorithms will look into site access to score the likelihood of a miRNA-mRNA interaction occurring.

- **Abundance of between miRNA target binding sites** - Abundance of 3' UTR miRNA binding regions. Many 3' UTRs can contain multiple miRNA binding sites [63]. The greater the number of binding sites, the higher the likelihood of off-shoot interactions, and actually lowers the likelihood of strong miRNA-mRNA interactions.

- **AU nucleotide flanking** - The higher the level of AU nucleotides flanking the seed site on the miRNA, the higher the likelihood of site depletion [64, 65].

- **GU wobble** - GU wobble is the term which refers to the number of G-U mismatches allowed in miRNA-mRNA interactions [5].
- **CDS, 5'UTR and 3'UTR restrictions** - Most miRNA-mRNA interactions occur in the context of the 3'UTR of the targetted mRNA. However, some algorithms also look into the whole coding sequencing region of the mRNA, and may also look into the 5' UTR, as well as the 3' UTR [66].
- **Size and position of miRNA-mRNA** - Size of interacting miRNA/ mRNA and position of interaction can also be taken into consideration [65, 67]. Length of the miRNA seed site can affect the likelihood of a miRNA-mRNA interaction occurring. Seed length can be between 6 and 8 nt long. Generally, 8nt (8mer) long seed seed sites show the best specificity and 6nt (6mer) long seed sites have poorer specificity for mRNA target sites [68]. Some algorithms take this information into account and score longer seed sites more preferably.

miRNA target databases

There are a number of miRNA prediction softwares available, some with experimentally validated miRNA-mRNA interactions, and others which are entirely algorithm based, and the features of the algorithms have been described above.

Experimentally validated databases include:

- ComiRNet [69]
- miRecords [70]
- miRSel [71]
- miRTarBase [72]
- miRWalk [73]
- MtiBase [74]

- starBase [75]
- Tarbase [76]

Interactions which have been functionally validated can be very helpful, as they can be used to record citations of miRNA work and provide confidence for interactions. The confidence provided by functionally validated interactions could also be extended to cross-species functional validation i.e. validated and peer-reviewed interactions from human work can aid mouse research. However, there are some caveats when working with such databases. These databases may include functional characterisation from a range of experimental techniques, including more reliable techniques like luciferase assays, and less reliable techniques such as RNAseq analysis. Deciphering between the reliability of the experimental techniques is an important factor to consider. Also, the quality of the database relies on the maintenance staff which curate them. This factor makes some of these databases (miRsel, Starbase, miRecords) much smaller than well maintained and updated databases (miRTarBase, miRWalk, Tarbase). Further differences in the stats of the databases are described in the following review [77].

Algorithmic prediction based databases include:

- HOCTAR [78]
- miRDB [79]
- microPIR [80]
- multiMiR [81]
- Pharmaco-miR [82]
- Targetscan [83]

Again, as with the experimentally validated databases, more regularly updated prediction based databases have a greater number of miRNA-mRNA interactions, and this can be seen in this review [77]. However, most of the interactions within these databases will not

have been validated, and they are predicted based on a number of features. Databases with alternate algorithms will lead to alternate predictions [59].

1.1.5 Importance of miRNAs

Over 1900 miRNAs, including over 500 unique miRNAs, have been identified in humans and these can be separated into 87 distinct families of miRNAs which are evolutionarily conserved genes that regulate over 60% of humans protein coding genes [60, 84]. This implies there is a positive selection for miRNAs to not undergo mutation and interestingly, coding DNA sequences have been speculated to be under a negative selection pressure to avoid being complementary to seed sequences of miRNAs. This indicates the important roles miRNAs have in biological processes [85]. The importance of miRNAs is evident as they are found in all cell types in mammals and play important roles in development and homeostasis [86, 87, 88]. However, more research is needed to identify how miRNAs regulate biological processes. The role of the miRNAs within the context of different biological niches has been under investigation for the last few decades; reviews can be found in bone formation, kidney homeostasis and cartilage formation to name a few [89, 90, 91].

There are several potential uses of miRNAs in the advancement of biomedical research. Firstly, several miRNAs have been reported to contribute to disease states in humans, so investigating their specific roles in these circumstances could be interesting. *miR-15* and *miR-16* have been reported to be deleted in many cases of chronic lymphocytic leukaemia [92]. Epstein Barr virus induces *miR-155* overexpression in B-cells by promoting survival factors and this contributes to B-cell lymphoma formation [93]. Both *miR-103* and *miR-107* are upregulated in obese/ diabetic mice and their silencing positively affects glucose homeostasis [94]. *miR-133b* is expressed at deficient levels in mid-brain dopaminergic neurons from Parkinson's disease samples [95]. Each of these miRNAs mentioned above are potential drug targets to explore for their respective diseases, and could also be classified as novel biomarkers within patients.

Secondly, miRNAs can be shuttled around biofluids [96]. This is unique and could provide medical researchers with a new diagnostic tool and I will explain why. There are

five methods by which miRNAs can be transported in biofluids: with high-density lipoproteins, complexed with an AGO2 protein, packaged in exosomes, packaged in microvesicles or packaged within apoptotic bodies and these methods have been reviewed and documented in *Kumar et al (2017)* [58]. Circulating miRNAs have been found at high expression levels within serum, blood, urine and other fluid, and they can be measured in fluids as a non-invasive diagnostics tool to classify patient health [97, 98]. For example *miR-141*, *miR-149*, *miR-299-5p* and *miR-135b* are found in the placenta of pregnant women [99]. Drug induced liver injury patients had *miR-122* and *miR-192* enriched in tissue and blood plasma, but the results found in the plasma are detected earlier than in the tissue, making miRNA plasma measurements a more valuable method of diagnosing liver injury [100]. *miR-21* and *miR-192* have been found in high quantities in the urine of liver fluke-associated cholangiocarcinoma patients, and post treatment, the amount of *miR-21* and *miR-192* decreased, providing potential non-invasive markers for a liver cancer and metrics to track cancer treatment [101].

Lastly, there is the potential of utilising the native functions of miRNAs to treat diseases. For example, *miR-9* reduces BRCA1 activity which could be a useful tool to treat BRCA1 mutant derived ovarian cancers [102]. Also, *miR-182* targets *BRCA1* in breast cancer cells and overexpression of *miR-182* leads to more irradiation and PARP inhibitor sensitive cells [103]. In some conditions, specific miRNAs are downregulated, and re-introducing them could revert disease phenotypes. For example *miR-140-5p* re-introduction in *in vitro* models of osteoarthritis (OA) leads to a reduction in expression levels of pro-inflammatory proteins such as NFkB1 and cartilage degradation proteins like ADAMTS5, which are contributing factors to painful and less functional joints in OA patients [104, 105, 106].

miRNAs are not yet viable for personal medicine

Unfortunately, the potential advantages of miRNAs for medical researchers and patients are not yet feasible. This is primarily due to the complexities of miRNA biology, which limits their potential use in biomedicine. The major hurdle being that a single miRNA can target many mRNAs and a single mRNA can be targetted by many miRNAs. Furthermore, different cell types and developmental stages are regulated by different miRNAs. This

makes understanding the role of a specific miRNA difficult. Also, some miRNA-mRNA interactions are redundant, so downregulation of a miRNA may not cause a noticeable phenotypic shift. To combat these complexities big data and computational approaches can be used, in synergy with wet-lab biology to better comprehend the complex biology of miRNAs.

1.2 Computational approaches to investigate miRNA-mRNA interactions

To best utilise computational techniques to better understand biological processes, high quality datasets are required, specifically longitudinal miRNA-mRNA expression datasets. The miRNAs and mRNAs should be measured at the same time points and ideally each time point should be measured multiple times. In this PhD, I analysed several longitudinal miRNA-mRNA datasets from a wide range of biological conditions including: chondrogenesis, kidney injury and HD. Here I will briefly introduce the datasets and analysis methods used to further efforts in computational miRNA investigation.

Longitudinal miRNA-mRNA expression datasets

Post-transcriptional regulation of mRNAs has been a popular focus of research for over two decades. Next generation sequencing techniques such as RNAseq, microarrays and microRNAseq have been used to generate miRNA-mRNA expression datasets. However, static comparisons between different conditions only provides a snap shot of the transcriptome at a given time. To capture more information about biological systems, longitudinal miRNA-mRNA expression datasets are created. Advantages of using this type of data includes: allowing for deeper analysis, providing information for kinetic models, and potentially highlighting oscillations and other temporal patterns. The downside is the rarity of finding high quality datasets due to their expense and labour intensive design.

These datasets are often stored in public repositories such as gene expression omnibus (GEO) or ArrayExpress [107, 108]. I will present three separate method for analysing

longitudinal miRNA-mRNA datasets, using the datasets listed below. The three methods are: bioinformatic analysis, kinetic modelling and machine learning.

- FA (Folic Acid) induced kidney Injury dataset - Bioinformatic Analysis [109, 110].
- UUO (Unilateral Ureter Obstruction) induced kidney Injury dataset - Bioinformatic Analysis [109, 110].
- Breast Cancer dataset - Bioinformatics Analysis [111].
- Hypoxic Breast Cancer dataset - Bioinformatics Analysis [112].
- Chondrogenesis dataset - Kinetic modelling [113].
- HD dataset - Machine Learning [114, 115].

Bioinformatic analysis

Standard bioinformatics analysis approaches of NGS data will often utilise differential expression (DE) analysis. This informs us of which genes are under the most change between two conditions. For longitudinal data there are specific methods which can be used, e.g. using the zero time point as the denominator, using a cubic spline to fit data, or analysing all time points at once. Depending on the length of the time course and the regularity of sampling, particular methods are more suitable. For example, the chondrogenesis dataset used in this thesis (Chapter (Ch)3 and Ch4) is six time points long, and most time sampling is irregularly measured within a range of 0 - 14 days (D). In this case, using the zero time point at the denominator is the best method for DE analysis [116]. After this, significantly differentially expressed genes (SDEGs) can be put through ontology or pathway enrichment to find biological significance. This method will identify gene functions or signalling pathways which are important in the dataset. These techniques have become standard tools for computational biologists. However, they are poor tools to identify specificity and untangle complexity. As such, they are useful upstream methods to inform downstream tools that are able to find specificity from large complex datasets. This concept is addressed as DE results are the input for a novel *R/Bioconductor* package called *TimiRGeN*, which was created in this PhD to integrate, analyse and produce small networks

from longitudinal miRNA-mRNA expression datasets [117]. Furthermore, *TimiRGeN* has several methods for longitudinal miRNA-mRNA pair analysis such as cross-correlation, regression and calculation of an odds-ratio. These are explained in Ch2, however I wish to highlight, this is the first miRNA analysis tool which helps to reduce the high volume of data found in a longitudinal miRNA-mRNA dataset. This is an important concept when aiming for hypothesis generation from big data. Several hypotheses from analysis with the *TimiRGeN R* package have been made. In Ch2 and Ch3 Gene regulatory networks (GRNs) have been constructed based on these hypotheses.

GRNs and kinetic modelling

GRNs are used in systems biology to display genetic interactions, and ultimately aim to make complex events in a signalling pathway readable. GRNs are used by: modellers as blueprints of kinetic models, information sources for bioinformaticians and as repositories for experimental design for wet-lab experimentalists. In collaborative work, GRNs can be seen as a meeting point between bioinformatics, modelling and experimental work. The GRNs topology can be altered as species can be added or removed, drug targets can be decided and the complexities of a system can be explained clearly. GRNs also serve as a means to address complexities in biological systems. For example, to represent feed-back loops, incoherent activity of genes, post-translational modifications, effects of drugs and displaying multiomic regulation. In this project, the ability of a GRN to clearly visualise multiomic regulation and complex interactions involving multiple miRNAs is used to generate a blueprint for a chondrogenesis model. A simple example of a GRN is shown in Figure 1.3.

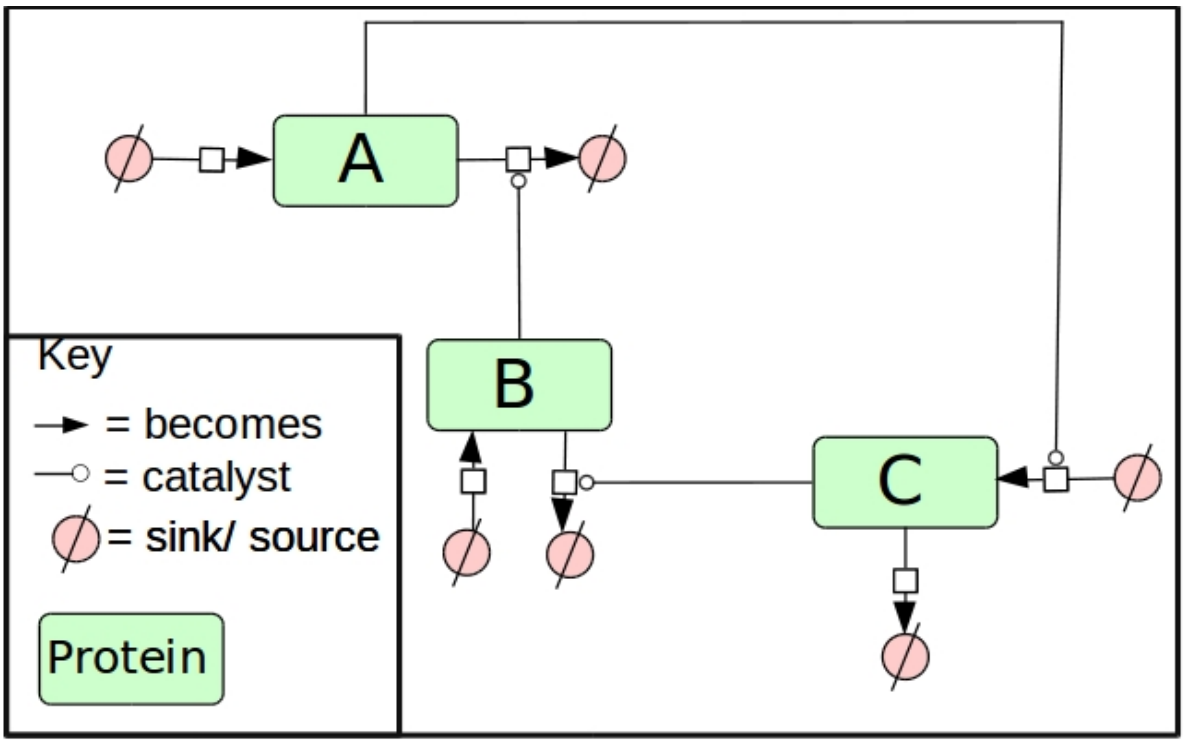


Figure 1.3: **Simple GRN with three species.** This is an example GRN which consists of proteins A, B, C which form a feedback loop. Here A catalyses the activation of C, B catalyses the degradation of A and C catalyses the degradation of B.

Due to the complex nature of many biological systems, solely using experimental work is labour intensive, expensive and difficult. To complement experimental work, we can deploy kinetic modelling, a systems biology technique to simulate biological events. This aids in creating a more cost-effective and efficient investigative process. Kinetic models can make predictions if sufficient data is available.

Kinetic models use GRNs as a blueprint to help establish key interactions and a suitable topology. To quantify a kinetic model, the behaviours of all the species must act as predicted. Two sets of data are required to establish if a model is usable and can make useful predications and both datasets should ideally be time series. A calibration dataset which is used to quantify model parameters. This is followed by an independent validation dataset. Here experimental perturbations are introduced and the model attempts to simulate the experimental results. Once this is accomplished, a fully validated kinetic model can now be used to make predictions. For example, modulation of multiple miRNAs is

difficult *in vitro*, and so instead simulations can be performed to predict the outcome of multiple miRNA modulations. Moreover, a well calibrated and validated model can be used to make theoretical predictions which may be unavailable in wet-lab studies due to technological limitations. Finally, potential drug targets can be tested *in silico*.

Biological process	ODE
source creates A	$dt/d = k1$
A degrades into sink	$dt/d = A * k2$
A catalyses C	$dt/d = A * k3$
source creates B	$dt/d = k4$
B degrades into sink	$dt/d = B * k5$
B catalyses degradation of A	$dt/d = B * A * k6$
source creates C	$dt/d = k7$
C degrades into sink	$dt/d = C * k8$
C catalyses degradation of B	$dt/d = C * B * k9$

Table 1.1: **Example ODEs for the simple GRN.** Based on the GRN shown in Figure 1.3, example ODEs are displayed in a table alongside the biological events they represent. Most parameters are kn. All functions are based on mass-action or constant flux for ease.

The kinetic models I create in this PhD are ordinary differential equation (ODE) based. These are used to convert biological processes into mathematical equations. For example, the simple model seen in Figure 1.3 can be represented by ODEs. A variety of functions and parameters can be used to inform the ODEs, and the aim is to generate simulations of biological behaviour which match the calibration and validation experimental data. For that reason, only longitudinal data is suitable for kinetic model creation. Table 1.1 informs how the simple model presented in Figure 1.3 could be modelled based on simple mass-action ODEs [118]. Mass-action is used as the default function in COPASI (systems modelling software) because of its simplicity [119]. A species (A, B or C in the simple example) is modulated by a parameter. The larger the parameter, the greater the affect on the rate of the species. Constant flux is used when a Protein is being created from source. This

means that protein will be inputted in the model at a rate determined by kn. Multiple GRNs and a fully validated multi-miRNA chondrogenesis model is presented in Ch4.

Machine learning Finally, machine learning techniques are becoming prominent in biosciences, and its application on big multiomic data sets could lead to finding novel patterns, biomarkers or drug targets. Machine learning requires splitting a data set into training and testing data. Algorithms are trained on the training data, and then applied to predict or classify features in the testing data. The complexity of large longitudinal multiomic datasets makes the potential uses of machine learning a novel and useful tool. A ML project is described in Ch5.

1.3 Contributions from this PhD

I have used three broad computational techniques to utilise information from longitudinal miRNA-mRNA datasets. These techniques are: big data bioinformatics via the development of the *TimiRGeN R* package, kinetic modelling and machine learning. Below I briefly describe the rationale and output of each of these projects.

TimiRGeN R package

I developed the *TimiRGeN R/ Bioconductor* package. This is a novel tool for integration, analysis and network generation of longitudinal miRNA-mRNA datasets. This tool can help researchers make sense of their miRNA-mRNA expression data and find miRNA-mRNA interactions within signalling pathways of interest. From here, a network representation of the filtered miRNA-mRNA interactions can be created and miRNA-mRNA interactions can be analysed using a suite of methods including cross-correlation, regression and clustering, or the results can be exported into *PathVisio* or *Cytoscape* [120, 121]. This type of open ended analysis can help users to identify how the filtered miRNA-mRNA interactions may be regulating the signalling pathway of choice, thus aiding in hypothesis generation. Hypotheses generated like this can be formalised with the creation of GRNs. Figure

1.5 shows how *TimiRGeN* has the potential to become a part of any miRNA-mRNA expression analysis project. Ch2 goes over the *TimiRGeN R* package, its functions, some results including figures generated by the package and GRNs constructed with its aid, and describes the steps needed to successfully create an *R/Bioconductor* package. This package lead to a first author publication in *Bioinformatics*; presented in Appendix A [117].

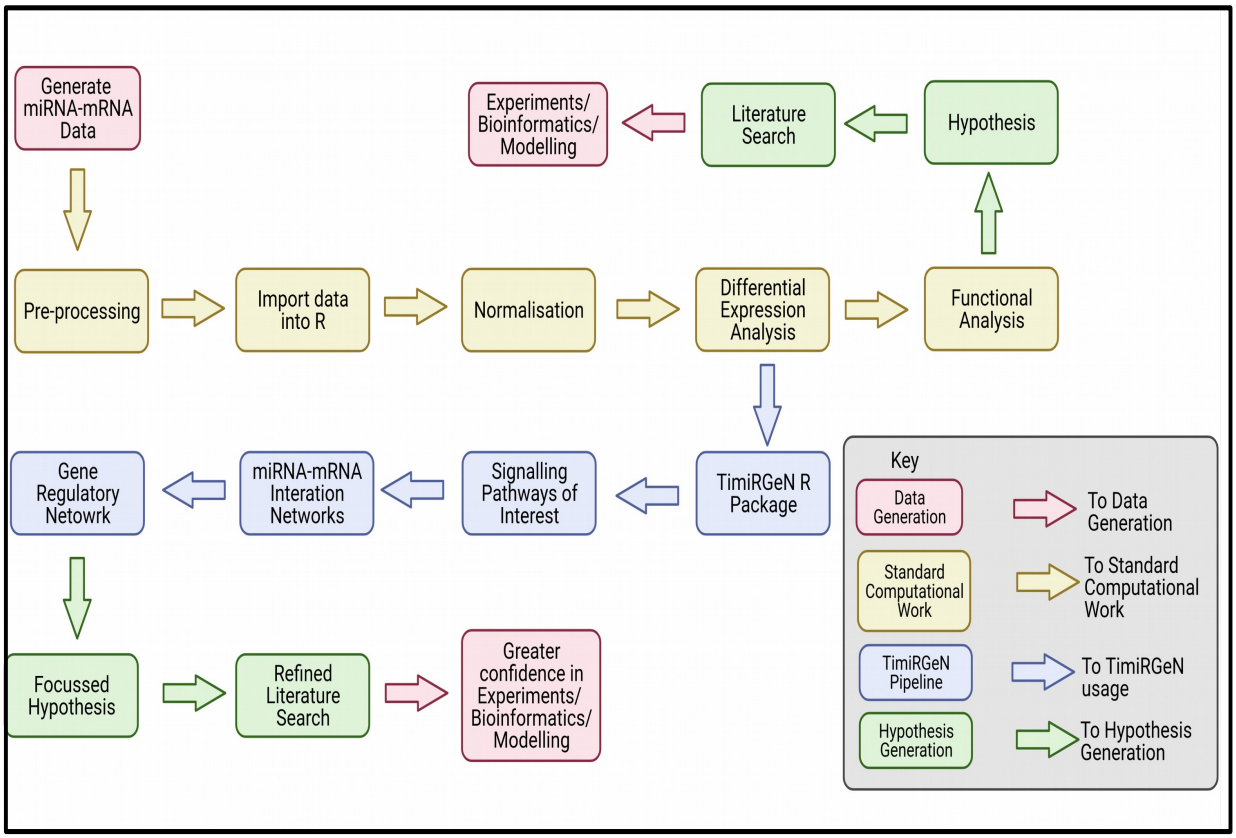


Figure 1.4: ***TimiRGeN* as a part miRNA-mRNA data analysis.** This schematic shows the standard computational process (yellow) to generate hypothesis (green) from miRNA-mRNA expression data (red). *TimiRGeN* analysis (green) provides an alternate path of analysis.

Gene regulatory networks found by *TimiRGeN*

GRNs have been generated from processing data with *TimiRGeN*. Several longitudinal miRNA-mRNA datasets have been analysed and GRNs have been generated from a drug induced reversible mouse kidney injury model datasets and a chondrogenesis dataset

[109, 110, 113]. These GRNs formalise hypotheses generated by the results from *TimiRGeN*. A novel feature of generating GRNs from network based approaches, rather than solely from literature, is the possibility of finding of indirect influences not reported in the literature. Both Ch2 and Ch3 showcase the GRNs generated from analysing longitudinal miRNA-mRNA datasets with *TimiRGeN*.

Validated *miR-199b-5p* chondrogenesis model

Ch3 also introduces chondrogenesis and the importance of miRNA regulators in this process. Using *TimiRGeN*, a chondrogenesis dataset generated by our collaborators is analysed [113]. Functional analysis by the tool identified the TGF-beta signalling pathway occurring during several time points. Further *in silico* analysis finds *hsa-miR-199b-5p* to be the second most positively changing miRNA within the TGF-beta signalling pathway; only *hsa-miR-140-5p*, which was mentioned is subsection 1.1.5, had a greater positive change over the time course. In contrast to *hsa-miR-140-5p*, *hsa-miR-199b-5p*'s regulation of chondrogenesis is relatively unknown, and only one recent paper reports on this [122]. During the *in silico* investigation, *CAV1* was identified as a mRNA target for *hsa-miR-199b-5p* and a *hsa-miR-199b-5p* homologue, *hsa-miR-199a-5p*. These miRNA-mRNA interactions became the foundation of an extensive literature search, GRN construction and kinetic modelling project which is continued in Ch4.

GRNs are created to explain how *miR-199a-5p* and *miR-199b-5p* may be regulating chondrogenesis via *CAV1*. This highlighted a well researched chondrogenesis signalling pathway, which is the RHoA/ROCK1 pathway. TGFB3 and CAV1 are upstream of RHoA/ROCK1 and chondrogenic biomarkers are downstream. A GRN was constructed to display the story of how *miR-199a/b-5p* may regulate chondrogenesis. Validation data from collaborators in the Young lab identified inhibition of *hsa-miR-199b-5p*, lead to anti-chondrogenic changes. A kinetic model was created to capture the complexities of this system and make predictions such as: the indirect affect on *hsa-miR-140-5p* and how *hsa-miR-199a-5p* inhibition will affect the modelled system. This work has lead to further experimental and modelling work which is mentioned in Ch6.

Detecting predisposition to Juvenile onset HD

I used ML to identify a set of genes within the context of juvenile onset Huntington's disease (JHD). The longitudinal miRNA-mRNA dataset used had mice that were sacraficed as different ages: 2M (month), 6M and 10M [114, 115]. To predict predisposition, the older mice (6M and 10M) underwent DE analysis to identify common SDEGs (significantly differentially expressed genes). Selected genes are used as the features for learning. Older mouse samples are treated as the training dataset and younger mouse samples (2M) are treated as the testing dataset. 15 different classifiers are used on the tested on the training and testing datasets. Logistic regression performed the best, and the resulting model detects WT samples with 100% accuracy, and the HD samples with 74% accuracy, thus further work is needed. Ch5 presents this work and Ch6 discusses further work that will be completed after the PhD.

CHAPTER 2

TIMIRGEN R PACKAGE

2.1 Background

This PhD focuses on using data science approaches to investigate longitudinal miRNA-mRNA expression datasets. From the literature/ tool search presented in Figure 2.1/ Table 2.1, no current bioinformatic tool is available to analyse longitudinal miRNA-mRNA expression datasets. To fill this niche, the *TimiRGeN R* package was developed to integrate, perform functional analysis and to generate small networks from longitudinal miRNA-mRNA expression datasets. This tool was inspired by many of the tools reviewed in Ch1, however it utilises many unique features which makes it a more versatile tool. In this section I will detail the key features of *TimiRGeN* which makes it a useful new tool for the computational biology community. Within my own research, this tool was used as a means to bridge large longitudinal miRNA-mRNA datasets and hypothesis generation, aid in GRN construction and longitudinal data analysis.

2.1.1 Comparison of miRNA-mRNA integration and analysis tools

The proceeding methods described above are individually useful on their own, but have potential to be even more powerful if they are used in combination with one another. As a part of this PhD I will investigate computational methods to design GRNs from big data

bioinformatics approaches. These techniques could contribute to one another. Big data bioinformatics provides global analysis of a large amount of data but has poor power when it comes to hypothesis generation and specificity. Whereas, GRNs have great specificity as they can represent small biological systems. Classical GRN construction is primarily literature driven. Large datasets, provided a more objective means for GRN construction. However, there is a major bottle-neck to overcome, which is how to reduce the sheer volume of data from large datasets to generate hypothesis which can be represented by GRNs. To add onto this complexity, in this PhD I will specifically be working with longitudinal miRNA-mRNA datasets.

We require a tool which can reduce the volume of data from found in longitudinal miRNA-mRNA datasets so GRNs can be more easily constructed. It would also be desirable for a tool which can use curated signalling pathways because mechanistic information from these is useful when constructing GRNs. This is as many signalling pathways (e.g. KEGG, Wikipathways) utilise literature and experimental work when constructing pathways [123, 124]. To establish which tools are available miRNA-mRNA integration and data reduction, a total of eleven tools are reviewed. These tools can be catagorised by their sources, which are: *Bioconductor*, *SourceForge*, web-based and locally installed software.

Bioconductor

Bioconductor is the largest repository of *R* packages for biological data exploration [125]. A tool which is accepted and maintained in *Bioconductor* is one which must contribute to the field of computational biology. For this reason, *Bioconductor* has very strict criteria for package induction. Several tools within this repository have been made to help researchers better understand miRNA-mRNA expression data sets, however none of the miRNA-mRNA integration tools which I tested could effectively analyse longitudinal datasets.

For example, *miRIntegrator* is a tool which can reduce the sheer volume of data, to the point where GRN generation is feasible, however it cannot analyse longitudinal datasets, only works with human data and many of its network generation abilities have been poorly maintained [126]. An inability to analyse longitudinal datasets is a common aspect of

other notable packages, such as *anamiR* [127]. This package generates a large matrix of miRNA-mRNA interactions and identifies which interactions occur in up to ten different miRNA-mRNA interactions databases. While this is a useful idea for exploration, using many databases like this has limitations. Different prediction databases use different methods to determine miRNA-mRNA interactions: seed sequence - 3' UTR complementary binding (TargetScan, miRDB), whole miRNA - 3' UTR complimentary binding (miRanda), thermodynamics (PicTar, DIANA) or hybrid methods (RNAhybrid) [79, 83, 128, 129, 130, 131]. Reviews comparing these different methods concluded that seed sequence - 3' UTR complementary binding approaches lead to the most true positives and the least number of false positives [132, 133]. Furthermore, many of these databases are not regularly updated. Unfortunately, *anamiR* also suffers from lack of regular maintenance and most of its features are currently not working.

Newer methods such as *spiderMiR* can reduce data to start generating hypothesis, but it does not cater to longitudinal datasets and its ability to utilise existing signalling networks is limited [134]. Considering the increase in the amount of longitudinal multiomic datasets being generated, it was surprising that *Bioconductor* did not yet have a tool to support this type of data. This motivated me to look outside of *Bioconductor*.

SourceForge

SourceForge is an open source program development platform which has miRNA based analysis tools. *miRComb* can integrate and analyse miRNA-mRNA expression data [135]. Similar to *anamiR*, it utilises multiple miRNA-mRNA prediction databases which use different prediction approaches. *miRComb* can generate a large matrix of miRNA-mRNA interactions and mine out commonly found interactions. This is a useful technique, however it is slow because *miRComb* assumes every miRNA can target every mRNA. *miRComb* had some capability to analyse longitudinal datasets, however it only uses the start and end time points, ignoring all intermediate time points, making its longitudinal analysis limiting. Despite some promise, *miRComb* is also not regularly maintained, and is not possible to use at the moment because of multiple bugs. Another tool for miRNA research on *SourceForge* is *sigterms* [136]. This tool identifies miRNA-mRNA interactions which may occur

between genes of interest and multiple databases (TargetScan, miRanda and PicTar). Unlike many other tools, *sigterms* distinguishes between databases which utilise different prediction algorithms. Unfortunately neither tool could analyse and integrate longitudinal miRNA-mRNA datasets.

Web-based

There are many web-based miRNA-mRNA integration and analysis tools, such as *miRNet*, *miRTarVis+*, *ToppmiR* and *MAGIA2* [137, 138, 139, 140]. All of these tools have extensive visualisation properties, however there are two main issues with them. Firstly, they do not handle longitudinal datasets, and secondly, most of these tools do not reduce the data enough to start formalising GRNs. A notable exception is *MAGIA2*, it uses a novel machine learning approach to produce small networks, from which GRN construction could be possible. However, the limitation of using *MAGIA2* for GRN construction is the lack of signalling pathway information it produces. To truly understand how a miRNA is influencing a target, downstream and upstream information is essential. Many web-based tools lack this information so none could be used to analyse our data for GRN construction.

Locally installed

Finally, some miRNA-mRNA integration tools can be locally installed. This includes *DREM2*, a java-based tool which identifies when miRNA-mRNA interactions are likely to be taking place along a longitudinal multiomic dataset [141]. Whilst this tool does handle longitudinal multiomic data, the output is lacking for hypothesis formalisation and GRN design. There is limited information on how the miRNAs may be affecting upstream/ downstream processes of their mRNA targets, and there are no network generation options. The last tool reviewed here is called *miARMa-seq*. This is a complete miRNA-mRNA expression dataset pre-processing, normalising and functional analysis pipeline [142]. This is a unique tool which could be useful, however, it does not support longitudinal analysis. Like many tools in this review, it can identify miRNA-mRNA interactions and perform functional analysis, but because it does not utilise signalling pathways or provide network outputs. So *miARMa-seq* cannot indicate how the miRNA-mRNA interactions may be influencing

the mechanistic pathways; making it hard to grasp what the miRNA-mRNA interactions are doing. Also, like many other tools, *miARMa-seq* does not reduce the sheer volume of data, making it difficult to begin GRN construction.

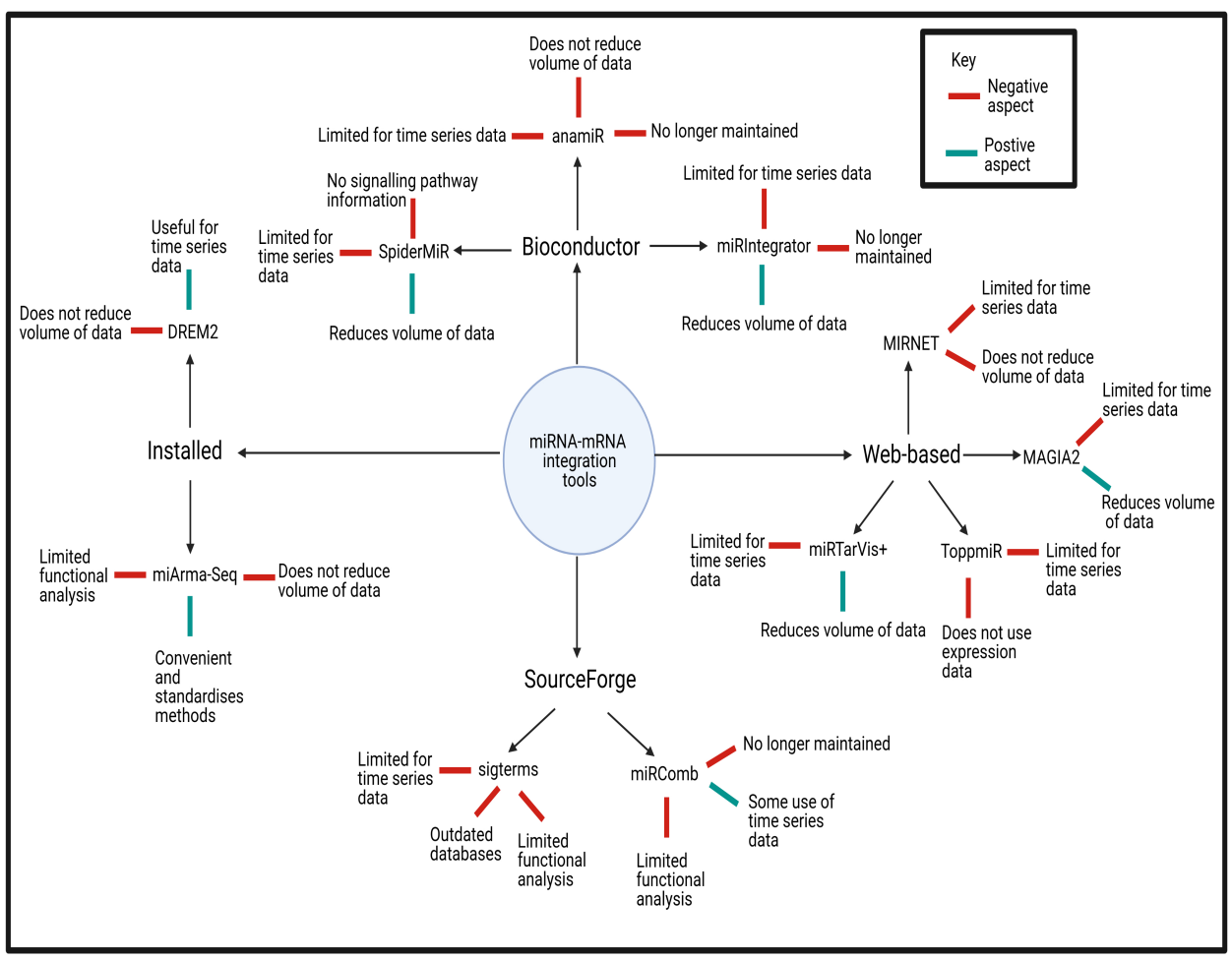


Figure 2.1: **Current miRNA-mRNA integration tools.** Mind map showing many of the current miRNA-mRNA integration tools. These tools were sorted as *Bioconductor*, Web-base, Installation and *SourceForge* tools. Each tool have some positive (blue) and negative (red) aspects labelled to them.

Tool name	Availability	Time	Funct analysis	Reduction	Updated
anamiR [127]	Bioconductor	×	✓:Kegg,Reactome,+	✓	2018
DREM2 [141]	Installation	✓	✓:GO	×	2020
MAGIA2 [140]	Web-based	×	✓:DAVID	✓	2012
miARMa-seq [142]	Installation	✓	✓:GO,Kegg	×	2019
miRComb [135]	SourceForge	✓	✓:GO,Kegg	✓	2020
miRIntegrator [126]	Bioconductor	×	✓:Kegg,Reactome	✓	2016
miRNet [137]	Web-based	×	✓:GO,Kegg	×	2021
miRTarVis+ [138]	Web-based	×	×	✓	2020
Sigterms [136]	SourceForge	×	✓ : GO	✓	2009
SpidermiR [134]	Bioconductor	×	×	✓	2020
ToppMiR [139]	Web-based	×	✓:GO	✓	2021

Table 2.1: **Comparison of miRNA-mRNA integration and analysis tools.** Several tools are *R* packages from *Bioconductor* or *SourceForge*, and others are either web-based or can be locally installed. Some can handle longitudinal data. Functional (Funct) analysis is often performed with GO, Kegg, Reactome, DAVID or others (+) and some of the tools are able to reduce the volume of data. The final column on this table shows the year of when each tool was last updated.

In conclusion, there are many miRNA-mRNA integration tools from a variety of different sources (Figure 2.1, Table 2.1). However, computational biology requires more sophisticated tools for increasingly more complex datasets. Now I introduce *TimiRGeN*. A novel *R/ Bioconductor* package, developed and maintained by myself which fills this analytical gap.

Target audience

This tool is aimed at experienced users of *R* and *Bioconductor* [125]. This tool is especially useful for researchers with longitudinal miRNA-mRNA datasets, however parts of the tool can also be used on static datasets. The tool has been presented in several inter-

national workshops and conferences, both orally and as poster presentations. Links to all presentations can be found in Appendix B.

2.2 Results

2.2.1 Features of *TimiRGeN*

Input Specifics

TimiRGeN is a flexible tool which is best used after DE analysis. It caters to a range of longitudinal DE outputs. *Spies et al (2019)* systematically contrasted multiple longitudinal DE methods, and concluded fewer false positives are found from a time course dataset with < 8 time points when using a pairwise DE method and fewer false positives are found from a dataset with $>= 8$ time points when using specific time series DE tools such as *MaSigPro* and *splineTC* [116, 143, 144].

DE is the principle processing step prior to *TimiRGeN* analysis. This makes the tool somewhat universal in its analysis, because no matter how the data is sourced (i.e. microarray, RNAseq, single cell Seq or dropSeq) or processed (i.e. DESeq2, limma, edgeR), DE has several standard outputs [145, 146, 147]. DE will produce confidence scores e.g. adjusted P value, P value, FDR, z-score, ect, and magnitude scores e.g. log2FC. If users wish to use a pairwise DE approach, they should extract one of both result types (confidence and magnitude) for each gene, from each DE analysis. Each gene should ideally be significantly differentially expressed in at least one of the DE analyses. However, *TimiRGeN* does have filtration options in-case non-significantly differentially expressed genes are within the input data. Users should also use a sensible common denominator. For example, the FA induced kidney injury dataset is analysed using pairwise DE [109, 110]. This dataset is an irregularly measure 14 day (D) time course with a zero time point, and the DE analyses performed were: D1/D0, D2/D0, D3/D0, D7/D0, D14/D0. Finally, all time points for the miRNAs and mRNAs should be the same. Again in the FA dataset, there was a D28 time point for the miRNAs, but this time point was not usable as input for *TimiRGeN* because the mRNA data did not have measurements for D28 [109, 110].

If users have longer datasets they may wish to use a variety of other options for DE, including: established longitudinal analysis tools such as *MaSigPro* or *SplineTC*. Furthermore, users can use more popular tools such as *DESeq2* which has a time series analysis method which uses the LRT (Likelihood ratio test) method to test all time points at the same time [143, 144, 145, 148]. The issue here is that each of these methods have different outputs, and fixing *TimiRGeN* to be able to process a few of these would make the tool inflexible. Thus, the burden of processing the input data from non-pairwise DE analysis methods to be "*TimiRGeN*-friendly" falls on the users. Once non-pairwise DE has been performed, users should filter out non-significantly differentially expressed genes from averaged counts or expression levels. From here, the filtered counts/ expression levels can be used as input for *TimiRGeN*. Again, the time points for the miRNAs and mRNAs need to be the same. Overall, *TimiRGeN* can handle data from non-pairwise DE methods, just as well from pairwise DE methods, though use of non-pairwise DE requires some alternative methods for data wrangling and this is explained in detail in subsection 2.2.5 where a breast cancer dataset with nine time points is analysed [111].

Overall, DE methods produce different outputs, so automating DE-*TimiRGeN* would be inflexible. Leaving the input to users means more flexibility, but also means users would need bioinformatics experience. To further add on flexibility, the tool works on data from microarrays and RNAseq, and can analyse datasets from multiple vertebrate model organisms, including humans, mouse, rat and zebrafish.

Analysis approaches

Unlike many other miRNA-mRNA integration and analysis tools, *TimiRGeN* gives users the option to analyse their miRNA and mRNA data combined or separately. The combined approach is by-far more useful. Knowledge of miRNA activity within signalling pathways is limited so miRNA specific functional analysis is limited. Though, the incorporation of miRNAs within WikiPathways is growing, and there are regular updates. In time, miRNA specific functional analysis may become easier. Furthermore, since miRNA or mRNA data can be analysed individually, a user can analyse non-multiomic datasets with *TimiRGeN*.

Functional analysis repositories

Functional analysis is a common objective of big data analysis. It is used to identify biological processes, signalling pathways, conditions and phenotypes which are statistically over-expressed within a set of genes. Common sources for functional analysis are GO terms, KEGG pathways and Reactome pathways [123, 149, 150]. However, *TimiRGeN* takes advantage of Wikipathways [124]. This is a community based repository of signalling and mechanistic pathways from a broad range of species. Wikipathways is a growing resource, which is updated on a monthly basis, and as pointed out previously, there are a growing number of miRNA related/ incorporating pathways found within Wikipathways [124]. The additional advantage of using Wikipathways is the cross-platform capabilities of Wikipathways, *PathVisio* and *Cytoscape*. This feature proved useful during the development of *TimiRGeN*. The disadvantage of only using WikiPathways is that we may be missing important signalling pathways that are present in other repositories such as KEGG and Reactome [123, 150]. One option is to use *OmniPath* instead, which is a collection of all mechanistic pathways [151]. Though this will mean forfeiting the cross-platform capabilities between Wikipathways and *Pathvisio*. Another disadvantage is that our functional analysis is reliant on publicly curated databases, which may be biased towards certain biological niches.

Gene IDs

In order to make the most of the cross-platform capabilities of Wikipathways, specific gene ID codes are needed. Most Wikipathways (not all) are curated with either entrezgene IDs or ensembl gene IDs. As such, *TimiRGeN* allows users to retrieve either entrezgene IDs or ensembl gene IDs for the miRNAs and mRNAs for further analysis. However, the complex nomenclature system of miRNAs makes these annotation types inefficient for miRNA annotation. Neither annotation type is sensitive to strand specificity so -3p or -5p miRNA strands are annotated with the same IDs. As such, *TimiRGeN* generates adjusted entrezgene IDs and ensembl gene IDs so mature transcripts can be treated as individual RNAs, even if they are transcribed from the same gene. For example, *hsa-miR-140-3p* and *hsa-miR-140-5p* share the entrezgene ID of 406932 and ensembl gene ID of ENSG00000208017. *TimiRGeN* will create altered IDs: 406932.1 and

ENSG00000208017.1 for *hsa-miR-140-3p* and 406932.2 and ENSG00000208017.2 for *hsa-miR-140-5p*. These adjusted IDs will only be used during plotting and exporting data so will not interfere with functional analysis or database mining.

Data filtration approaches by *TimiRGeN*

All miRNA-mRNA integration and analysis tools attempt to filter big multiomic datasets, and *TimiRGeN* uses three different levels of filtration: confidence levels, pathways of interest and by miRNA-mRNA interactions. Reducing the volume of data not only allows for easier hypothesis generation, but also makes analysis less computationally intensive and faster.

Filtering by confidence levels

If pairwise DE was used as input, the users data will include genes that are differentially expressed in at least one pairwise contrast. If the combined analysis mode (miRNAs and mRNAs analysed together), genes are ordered into nested dataframes within lists based on time point. Here each gene at each time point can be filtered for significance independent of each other time point. If the separated analysis mode (miRNAs and mRNAs analysed separately) is being used, then each miRNA or mRNA can be filtered for significance independent of each other time point and each gene type (miRNA or mRNA). This type of filtering relies on the inclusion of a confidence level from DE e.g. adjusted P values. If users use non pairwise based DE, then this step should be performed before importing data into *TimiRGeN*.

Filtering by pathways of interest

TimiRGeN offers two distinct methods for functional analysis, both using the *rWikiPathways* API to identify pathways of interest [124]. The first uses overrepresentation analysis (ORA). The number of genes found at each time point (after filtering by confidence levels) and each species specific Wikipathway are contrasted. The Fisher exact method generates P values, which can then inform which pathways are most enriched at each time

point. If separated analysis is performed, the miRNAs and mRNAs at each time point will be assessed individually. It is common not to find many enriched pathways with the miRNAs only, for reasons explained before.

Alternatively, there is a temporal cluster analysis approach which uses *Mfuzz* [152]. During use of the combined analysis mode, fuzzy clusters are created based on the common genes found between the species specific Wikipathways and different time points (after filtering by confidence levels). This analysis can identify temporal patterns. If separated analysis is performed, the miRNAs and mRNAs need to be assessed individually.

These methods will highlight several pathways of interest which can be further explored for miRNA-mRNA interactions which may be regulating the selected pathways, within the context of the biological niche.

Input data from non-pairwise DE can also be explored using pathway enrichment or fuzzy clustering. With this type of input, the entire set of genes can be analysed via ORA to find pathways which are enriched for the whole time course. Also, genes from non-pairwise DE can also be analysed with fuzzy clustering. Then genes which correlate well to the clusters can undergo pathway enrichment to identify which pathways are most enriched within each cluster.

Filtering by miRNA-mRNA interactions

Once a pathway of interest is identified, genes found both in the pathway and input mRNA data are filtered. Every miRNA is assumed to have the potential of interacting with the filtered mRNAs. Correlations are calculated between every potential miRNA-mRNA interaction. The default correlation method is Pearson, but Kendall and Spearman are also options. These correlations are created using longitudinal changes in a DE results type which represents magnitude e.g. log2FC or averaged counts/expression. *TimiRGeN* creates a large correlation matrix, including miRNAs, mRNAs, IDs and correlations. In addition to this, up to three miRNA-mRNA target databases be mined from and added to the matrix as columns. miRNA-mRNA interactions being present or not present in a database

is represented by the addition of a 1 or a 0 respectively. *TimiRGeN* only uses TargetScan, miRDB and miRTarBase (rationale explained later in this subsection), so potential miRNA-mRNA interactions can be scored between 0-3 [72, 79, 83].

Users can utilise the features of the matrix to filter for miRNA-mRNA interactions with more confidence. The number of target databases which interactions are found in and correlations can be mined to filter out low confidence interactions. The resulting filtered miRNA-mRNA interactions can be used to proceed onto network generation. Overall, with these filtering options, potentially hundreds-of-thousands of potential miRNA-mRNA interactions are reduced to a more manageable amount, and users have several parameters which they can adjust to make the filtration steps more stringent or relaxed. This level of data reduction is rare among miRNA-mRNA integration tools.

miRNA-mRNA target databases

The *TimiRGeN R* package will generate a large matrix of potential miRNA-mRNA interactions, and these must be filtered for predicted interactions. From the vast range of algorithms available, I selected two predictive target databases: TargetScan and miRDB [79, 83]. Both algorithms used used seed site - 3' UTR complementary binding and filtered for interactions with high evolutionary conservation (see subsection 1.1.4). However, they have some differences, TargetScan has strict requirements for acceptable seed sites of miRNAs. Seed sites have to follow a k-mer logic rule. True seed site - 3'UTR interactions have to be either an 8mer (target sequence matches positions 2-8 of the miRNA followed by an adenine), 7mer-1A (target sequence matches positions 2-7 of the miRNA followed by an adenine) or 7mer-m8 (target sequence matches positions 2-8). There are also some rulings for 6mer class miRNA-mRNA interactions e.g. exact seed site match, but 6mer interactions are classed as poorly conserved. In contrast, miRDB is less restrictive as it only asks for a 2-8 nucleotide seed site sequence match. TargetScan also looks into AU content and seed site positions. miRDB also takes into account accessibility and free energy in their SVM model. Overall these two algorithms have some overlap in their approach. Seed site - 3' UTR was kept in mind during researching potential databases to use in *TimiRGeN*, because comparative studies found them to identify to least number of

false positives [132, 133]

In addition to two TargetScan and miRDB, I included miRTarBase to provide functionally identified miRNA-mRNA interactions [72]. I removed interactions which were captured with ”weak” evidence e.g. next generation sequencing, so only keeping the results of more robust techniques e.g. PAR-clip, luciferase assay ect. In contrast to the prediction databases, functional databases are much smaller. The assumption is that these three databases can be used to filter the large miRNA-mRNA correlation matrices created by *TimiRGeN* for interactions which users can have a higher confidence in.

Network generation

TimiRGeN has three options for network generation: 1) plot the filtered miRNA-mRNA interactions in *R* using igraph, 2) export filtered interactions to *PathVisio*, 3) export interactions to *Cytoscape* by using the *RCy3* package [120, 153, 154]. This open-ended style is unique among the miRNA-mRNA integration software presented in Table 2.1.

Generating networks can identify miRNA-mRNA interactions of interest and direct users to more interesting pathways. However, if too many miRNA-mRNA interactions have been mined from the correlation matrix, visualisation in *R* will be difficult. Exporting to *Cytoscape* can be better. Version 3.7 of *Cytoscape* or newer must be already opened, and the *cytoscapeping()* function needs be used to establish connection between *R* and *Cytoscape* [121, 154]. The mined interactions will be sent to *Cytoscape* by using the *cytoMake* function of *TimiRGeN*. From here, a user will have better visualisation and access to *Cytoscape* apps.

Another limitation of internal *R* network generation is that the networks will not contain the upstream and downstream signalling information from the pathway of interest. Exporting network data to *Pathvisio* can resolve this [120]. At the moment, importing the data into *PathVisio* is not automated, but *TimiRGeN* can streamline this process. The *makeMapp* function will generate a file which lists the miRNAs predicted to influence the pathway of interest. All the mined miRNAs can be imported into *PathVisio* using the mapp app and

selecting the mapp file. The imported miRNAs must be moved manually, to interact with their predicted targets. The *makeDynamic* function will generate a file which contains chronological miRNA and mRNA DE values which represent magnitude of change e.g. log2FC. This file can be imported into *PathVisio* to colour code the changes over time in the pathway of interest. With this, the longitudinal changes in a miRNA integrated pathway of interest can be visualised. This type of visualisation is excellent for bottom-up GRN building (see section 2.2). A *PathVisio* guide has been created and is linked to in Appendix C.

Hierarchical clustering

Genes participating in the miRNA-mRNA interactions predicted by *TimiRGeN* can be clustered in a hierarchical manner. A dendrogram and an associated heatmap can be generated to display trends within the genes of interest. Individual genes within the clusters can also be plotted along a smooth spline.

miRNA-mRNA pair analysis

TimiRGeN has several metrics to analyse predicted miRNA-mRNA interacting pairs. These metrics mainly rely on correlation and regression methods which are widely used in longitudinal dataset analysis [155]. One such metric is cross-correlation analysis. This measures similarities between two time series, in this case the longitudinal changes in a miRNA and a mRNA. This method can also identify delays and periodicity if the time series is of sufficient length. Furthermore, it can be used as a metric to further filter of miRNA-mRNA interactions as one would expect a miRNA which negatively regulates a mRNA to have highly dissimilar temporal trends, which can be identified with cross-correlation [156, 157].

Two distinct regression methods can also be used to assess miRNA-mRNA pairs. Firstly a predictive regression analysis is performed between miRNA-mRNA interactions. A single gene (mRNA or miRNA) is selected and any combination of its predicted interacting partners are used to predict the expression of the selected gene over the time course.

This is a useful tool when a single mRNA is being targetted by multiple miRNAs or when a miRNA targets multiple mRNAs. The R^2 value and P value are plotted, and it is possible to infer synergy, competition or dominance between miRNA-mRNA interactions. However, further specifics between multiple miRNA-mRNA interactions are unavailable.

A simple regression analysis can also be performed between a single miRNA and mRNA. Here the regression coefficient is used to calculate odds-ratio (OR) and 95% CI (confidence intervals). The odd-ratio identifies the likelihood of one time course influencing the behaviour of another time course. Within the context of *TimiRGeN*, the OR score reflects the likelihood of a miRNA influencing a mRNA over a time course. The CI display a range where there is a 95% likelihood of the mean of the data being within the CI range, and the smaller the CI range, the greater the likelihood [158, 159].

2.2.2 Bioconductor

TimiRGeN has been accepted as a *Bioconductor* package, as of release version 3.12. This means the package was found to meet the strict criteria of becoming part of *Bioconductor*. Tools here are anticipated to be relevant for biological research, and the code which the package was written in was seen to be of a high standard. It took a total of 9 months (January - September) for the package to be accepted, during which time, under review process, hundreds of changes were made to the package. Several datasets have been analysed with the package and GRNs have been generated from hypotheses which *TimiRGeN* helped to find. Some are explained below to provide examples of analysing longitudinal datasets with the *TimiRGeN R* package. Links for reproducibility and installation code are found in Appendix C.

2.2.3 Pipeline of *TimiRGeN*

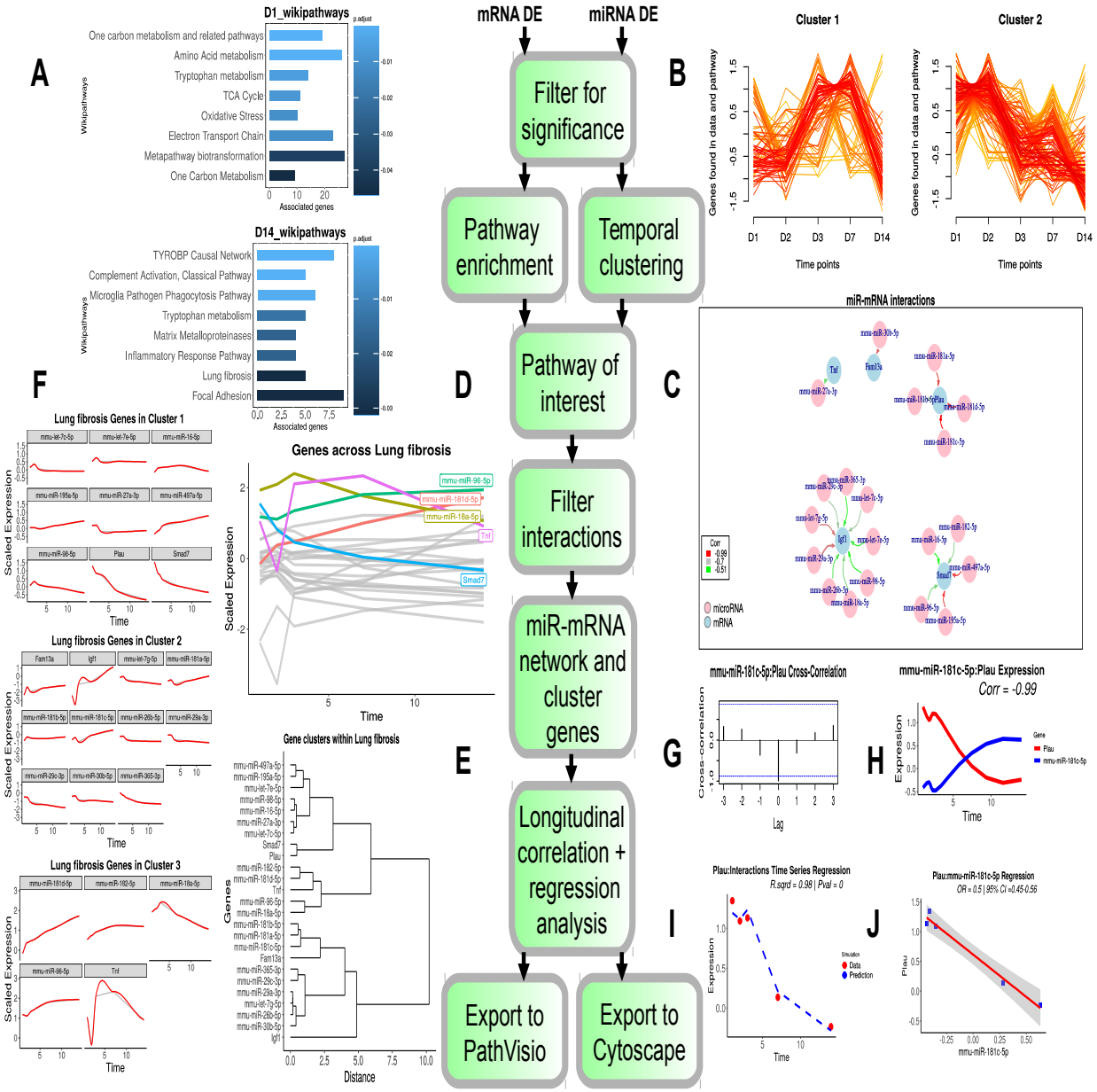


Figure 2.2: **Skeleton of the *TimiRGeN R* Package.** This pipeline is based on the combined method of *TimiRGeN* and it uses the FA dataset as a working example, thus it is based on RNAseq data and miRNAseq data which was undergone pairwise DE. The FA miRNA-mRNA data are input and filtered for SDEGs for each time point. Two method of functional analysis are available. **A)** time dependent pathway enrichment to identify enriched pathways at each time point. The enriched pathways are ranked in descending

order of adjusted P values. Only results from D1 and D14 are shown. Or **B)** temporal clustering which highlights global trends of pathways over the time course and clusters these trends. Two clusters are shown here. Each line represents a pathway and colour represents the fitness score a pathway has with a cluster. Ranking from lowest to highest are: purple, yellow, orange, red. After a pathway is selected for further analysis, miRNA-mRNA interactions within the selected pathway can be predicted by filtering for miRNA-mRNA interacting pairs using databases and correlation. **C)** Filtered miRNA-mRNA interactions can be viewed in *R*. Nodes are pink (miRNAs) or blue (mRNAs) and edges are colour coded by correlation over time. **D)** Behaviour of genes within the miRNA-mRNA interaction network can be viewed across the length of the time course and genes which pass a given threshold (greater than 1.5 within this example) are highlighted. **E)** Genes can also be clustered hierarchically for trend identification. **F)** Expression changes within the clusters also can be plotted and these line plots include a grey line (representing data points) and a red line (smooth spline over the data points). **G)** A selected miRNA-mRNA pair (e.g. *mmu-miR-181c-5p* and *Plau*) can be analysed with cross-correlation analysis. **H)** The selected miRNA (blue) and mRNA (red) can be displayed over the time course. The data can be scaled and interpolated over a spline and the correlation is displayed. **I)** Analysis with regression type methods can be performed on a selected mRNA or miRNA. *Plau* is selected here as an example. Its expression over time is predicted based on the selected miRNAs that are predicted to target it. In this example *mmu-miR-181c-5p* is selected to predict the time course trend of *Plau*. Log2FC values of *Plau* are displayed as red dots and the predicted Log2FC values of *Plau* is displayed as a dashed blue line. R^2 and P value are calculated and shown. **J)** Regression can also be performed between a single miRNA-mRNA pair. The OR between the two time series (miRNA and mRNA) can be calculated, along with the 95% CI. Correlation, P value, R^2 OR and CI are rounded to two decimal places. Network data can be exported to *Cytoscape* or *PathVisio*.

2.2.4 Combined miRNA-mRNA analysis with *TimiRGeN*

Here I present output of the *TimiRGeN* R package using a Kidney injury dataset.

Kidney injury dataset and data processing

To test the *TimiRGeN* R package, a ”gold standard” longitudinal miRNA-mRNA had to be used. The ideal dataset would have a minimum of three repeats per time point, and have miRNA and mRNA expression data taken at the same time points. There were a number of datasets that could have been used, and I settled on a mouse kidney injury dataset stored in a GEO repository. mRNA data was downloaded from GSE65267 and miRNA data was downloaded from GSE61328 [109, 110]. This dataset was suitable as the miRNA and mRNA time points matched, three biological repeats were taken at each time point, the time course consisted of six time points including a zero time point and the topic of kidney fibrosis was of interest in our research group.

FA induced kidney injury can be detected by miRNAs

FA injection is a method of simulating a chemically induced reversible acute kidney injury in model organisms such as mice [110]. In humans, acute kidney injury has a high mortality rate as around half of patients die [160]. The injury event triggers cell death which can cause loss of renal function. If the injury is mild the damage is reversible, but if serious, kidney injury can lead to decreased excretory release and contribute to clinical diagnoses such as CKD (chronic kidney disease) and renal failure [161]. A number of phenotypes can be characterised as hallmarks of CKD, including fibrosis, cell lysis and a heightened immune response. miRNA-mRNA interactions could be playing a role in contributing to CKD-like conditions. *miR-21a-5p* has been found in urine in CKD patients so it has been identified as a non-invasive biomarker, and this gene was highly upregulated over the time course of the data [162]. However *miR-21-5p* has also been identified as a non-invasive biomarker in several other conditions such as gastric cancer, bladder cancer, prostate cancer and lupus [163, 164, 165]. Thus, further investigation to find a panel of miRNAs which can be used to specify kidney injury would be a valuable asset to non-invasive biomarker research and personalised medicine. A further aid to this aim would be to generate GRNs to overlay multiomic data and present hypothesis for *in vitro* drug testing. In this subsection two GRNs have been generated to help explain the regulatory affects miRNAs have during FA induced reversible acute kidney injury.

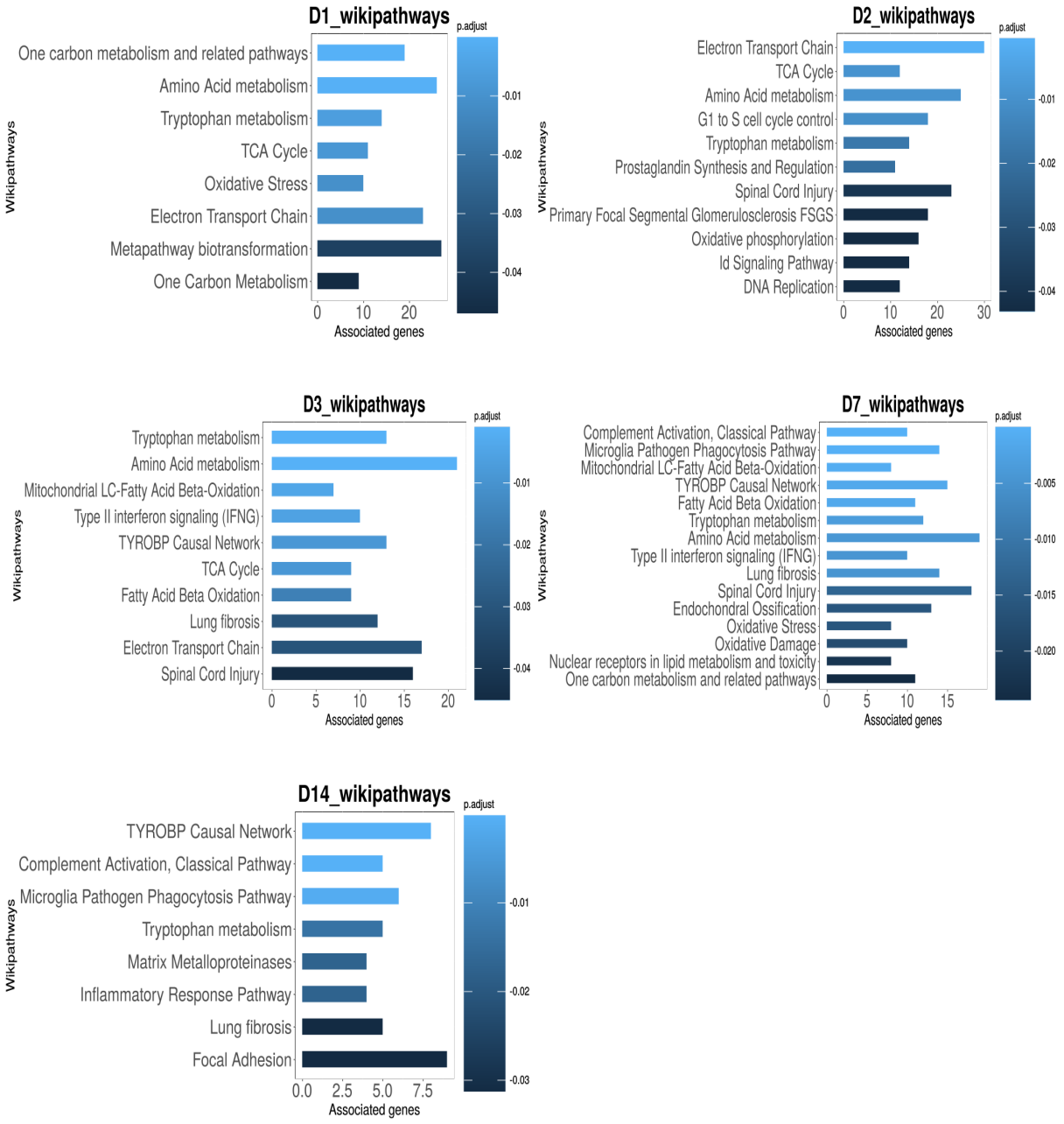


Figure 2.3: **Overrepresentation analysis bar plots created for each time point.** Pathways are ordered based on adjusted P values which are measured using the BH method. The darker the shading the higher the P value score. Barplots for D1, D2, D3, D7 and D14 are shown here. Plots have been displayed in a left to right chronological order.

IGF1 may act as a miRNA sponge

The FA miRNA and mRNA data were put through the combined method of analysis in *TimiRGeN*. Pathway enrichment found several signalling pathways which were consistently enriched throughout the time course. The Lung Fibrosis pathway (WP3632) was enriched at days: 3, 7 and 14, meaning it was a later acting pathway. As mentioned earlier, one of the weaknesses of using Wikipathways is that our functional analysis is bias towards curator input. As such, we would have expected a kidney related pathway to be enriched, but no such pathway existed at the time of the analysis. Thus, the Lung Fibrosis pathway is used with the assumption that the systems within this pathway are also found in kidney fibrosis.

Potential miRNA-mRNA interactions which regulate the Lung Fibrosis pathway were filtrated based on the following conditions: miRNA-mRNA interaction must have a Pearson correlation of less than -0.5 and must be found in at least two of the three target databases. 20 miRNA-mRNA interactions remained. Through further investigation with the *TimiRGeN R* package (detailed below), *IGF1* was found to as a potential miRNA sponge because our analysis predicted it to interact with nine miRNAs (Figure 2.4).

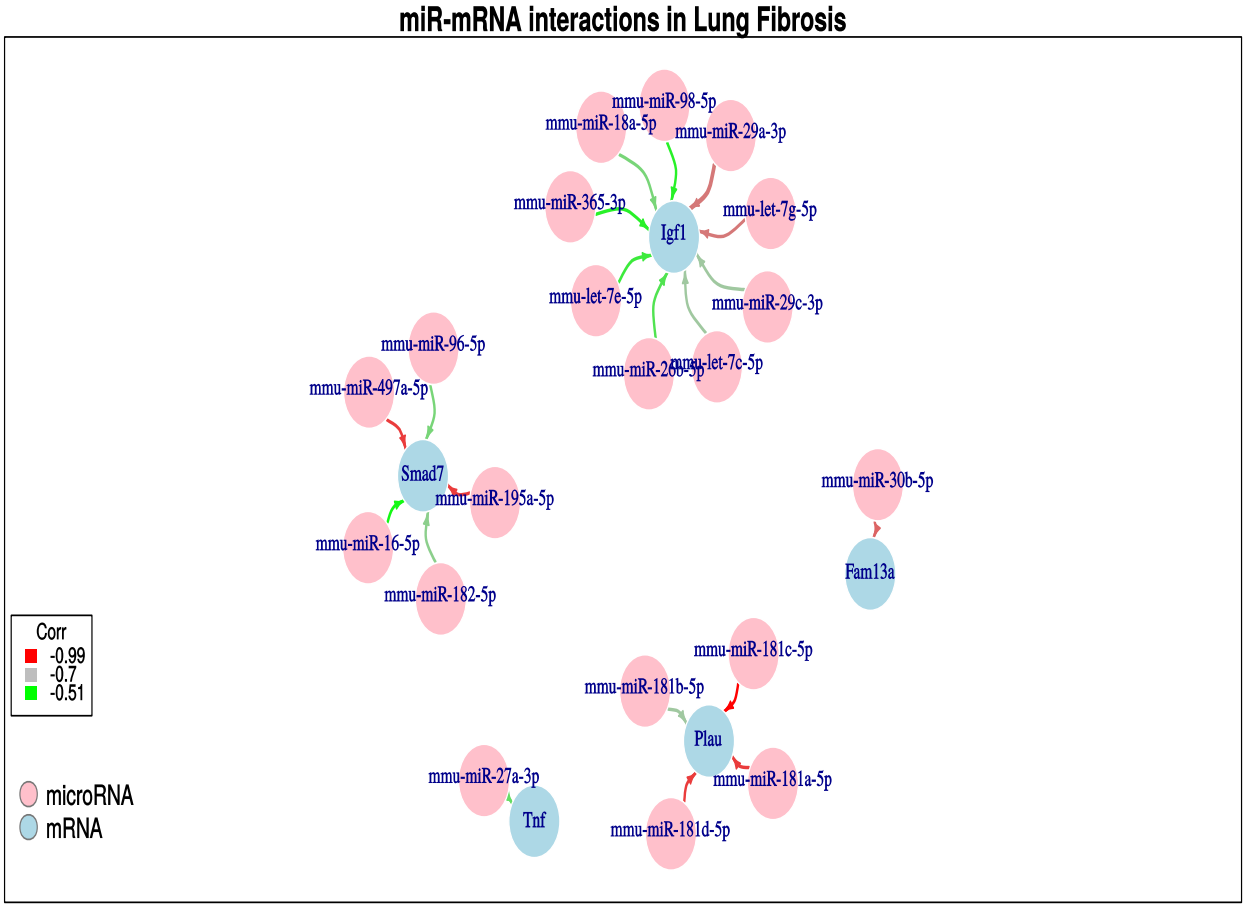


Figure 2.4: **igraph network displaying miRNA-mRNA interactions found after filtration.** mRNAs are blue and miRNAs are pink. Correlations (Pearson) inform the colour of the edges.

The miRNA-mRNA interactions could also be exported into *Cytoscape* for the option of using *Cytoscape* apps for further analysis. Figure 2.5 shows how the miRNA-mRNA interaction network looks in *Cytoscape*.

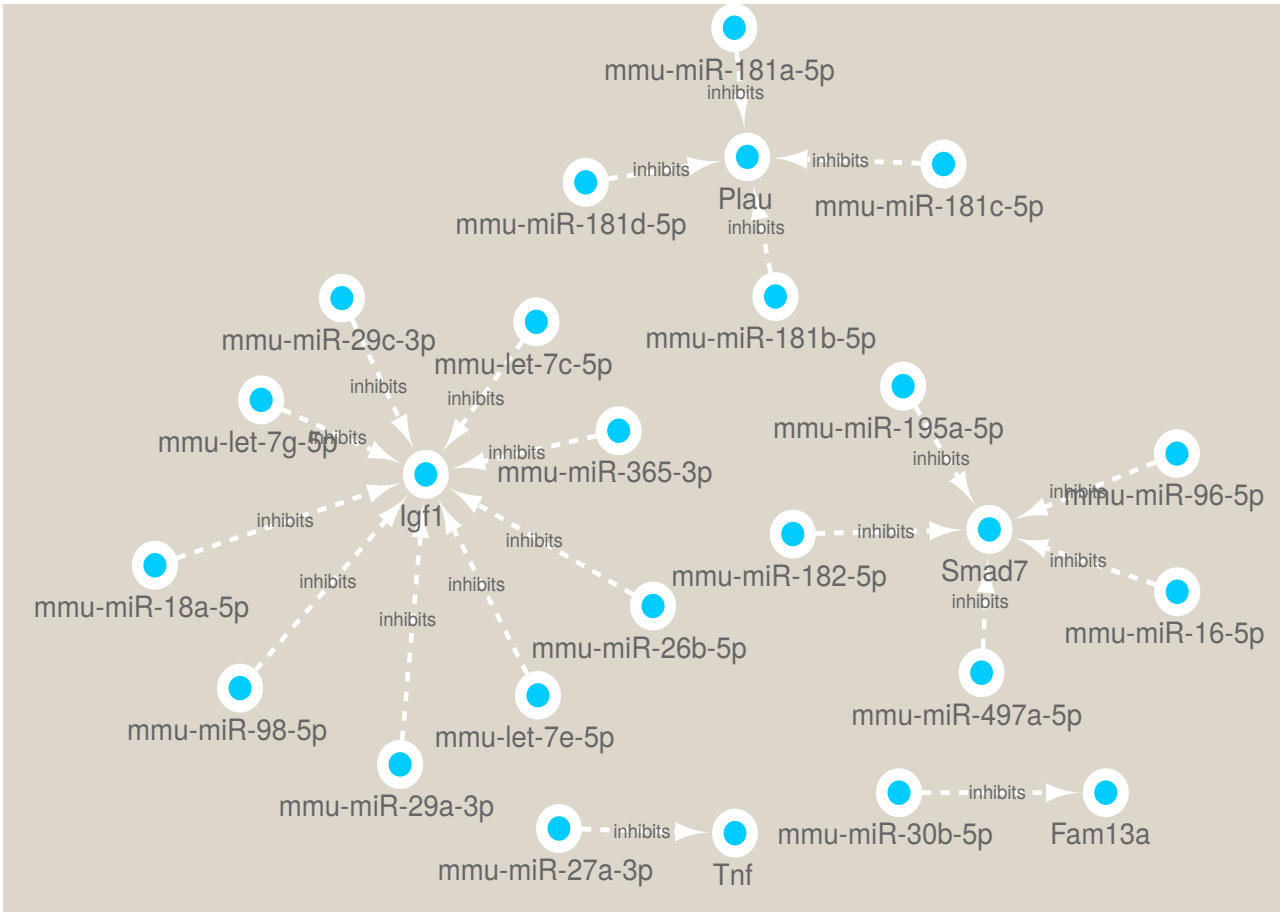


Figure 2.5: **Network data from *TimiRGeN* has been imported to *Cytoscape*.**

Instead, the results from Figure 2.4 were imported into *PathVisio*, along with dynamic information. This was in order to construct a miRNA integrated dynamic Lung Fibrosis pathway. A network like this can help in understanding the regulatory affects a miRNA may have on a signalling pathway. Moreover, this also allowed us to see how multiple miRNAs fit into a signalling pathway. Such as display is ideal for bottom-up GRN development.

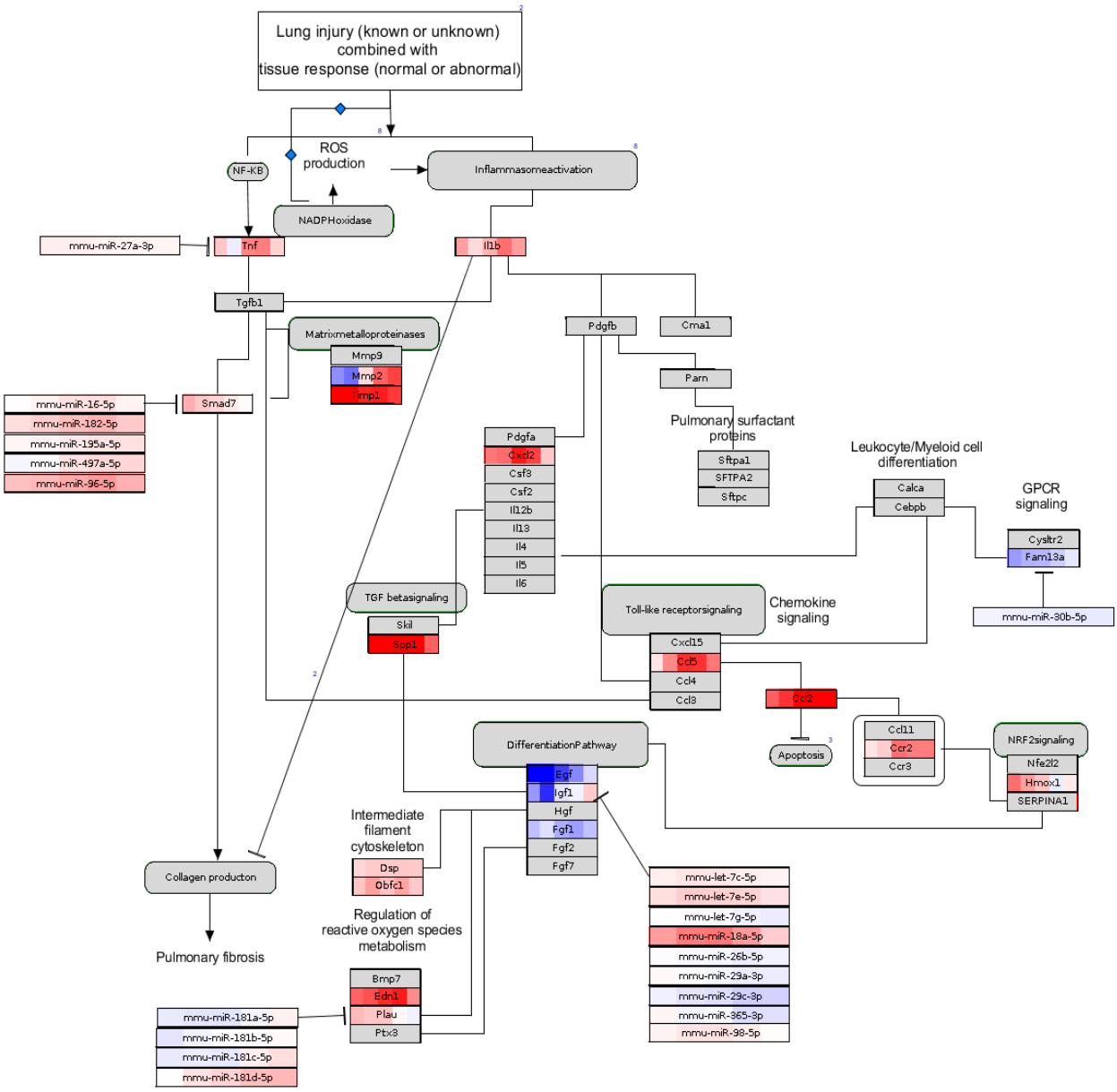


Figure 2.6: **Network data imported into *PathVisio*.** Each gene contains five segments, representing the five log2FC values from each time point in chronological order. Positive fold changes are red, negative fold changes are blue and little-no changes are white. Grey boxes are genes not found in the input data. miRNAs and dynamic information from *TimiRGeN* were added. Several sections of the original Lung Fibrosis Wikipathway were removed to make this display clearer.

Figure 2.4-2.6 highlighted the number of miRNAs interacting with IGF1. The Lung Fibrosis pathway in Figure 2.6 showed TGFB signalling is regulated by IGF1 and TNF and so

miRNA-mRNA interactions involving these genes were chosen for bottom-up GRN construction. The GRN in Figure 2.7 formalises the hypothesis that FA indirectly modulates multiple miRNAs which contribute to reducing expression of anti-fibrotic gene *Tnf* and increasing expression of pro-fibrotic *Tgfb* and *Igf1*. *TimiRGeN* analysis predicts *mmu-miR-27a-3p* targets *Tnfa1*, an antagonist of collagen promoting protein TGFB. Interestingly, *TimiRGeN* also predicts *Igf1* is a miRNA sponge for multiple miRNAs including miRNAs from the let-7 family (*mmu-let-7c-5p, mmu-let-7e-5p, mmu-let-7g-5p*), miR29 family (*mmu-mmu-miR-29a-3p, mmu-miR-29c-3p*), and other miRNAs: *mmu-miR-18a-5p, mmu-miR-26b-5p, mmu-miR-365-3p* and *mmu-miR-98-5p*.

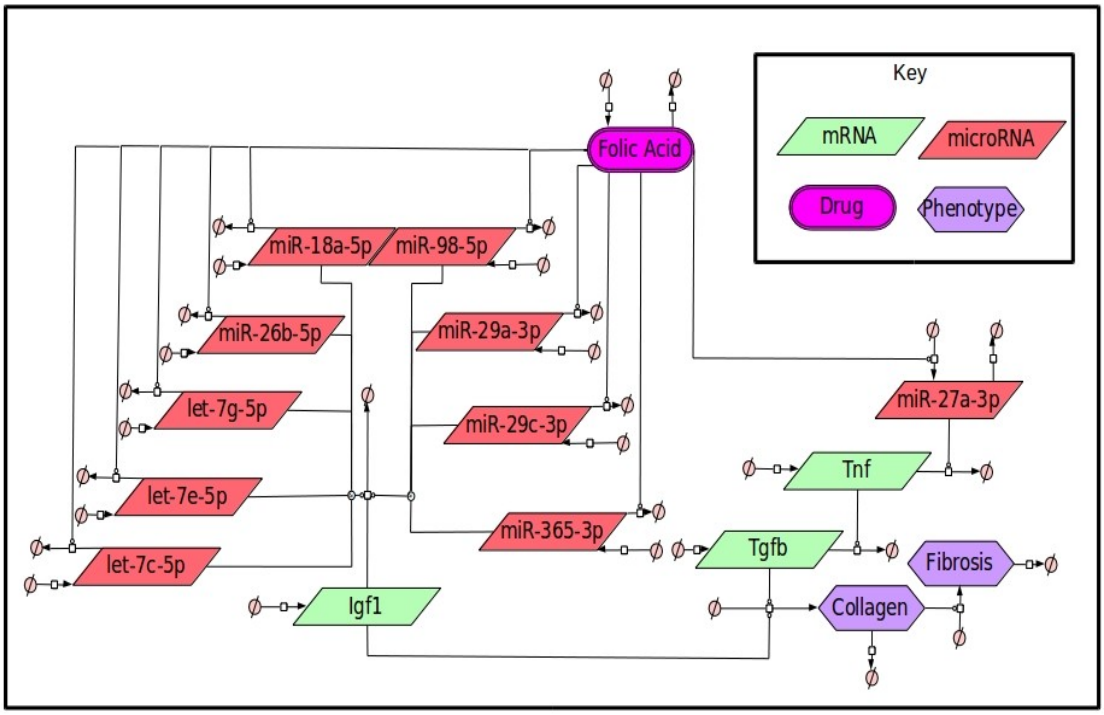


Figure 2.7: **GRN which displays *Igf1* as a miRNA sponge.**

The GRN presented above indicates *Igf1* is a miRNA sponge. Some of the predicted miRNA-*Igf1* interactions have been reported (*miR-18a , miR-26b, miR-98* and *miR-365*) [166, 167, 168, 169]. *let-7c-5p* has also been reported to target *Igf1*, and *TimiRGeN* predicted that other let-7 family genes including *mmu-let-7e-5p* and *mmu-let-7g-5p* also target *Igf1* [170]. Since let-7 family genes share most of their seed sequence, it increases the likelihood of other let-7 genes also targeting *Igf1* [171]. Finally, miR29 family members

have been predicted to target *Igf1*, and research indicates that *Igf1* can act as a ceRNA (competing endogenous RNA) for miR-29 family members [172]. ceRNAs are RNAs which regulate other RNAs by regulating miRNAs. Meaning, miRNAs sponged by *Igf1* are less able to regulate other target genes, leading to the target genes being upregulated [173]. It is unknown why *Igf1* may be a miRNA sponge, but *Igf1* is known to induce collagen production, which contributes to kidney fibrosis and CKD [174]. Exploration into the role of *Igf1* as a miRNA sponge in kidney injury conditions could be beneficial for therapeutics for CKD.

The GRN above has multiple assumptions for example direct/indirect negative regulation of the miRNAs targeting IGF1 by FA. In this case, the assumption was made because all of the miRNAs which targetted *Igf1* were decreasing over the time course, except *mmu-let-7e-5p* and *mmu-let-7g-5p* which peak around day 3 and then drop down in expression. Although, this assumption may be incorrect, the GRN can still be used as a Null hypothesis for further experimental or computational analysis.

The genes found to be involved in Figure 2.4 can be explored over the 14 day time course. In Figure 2.8, the genes which are higher or lower than a user defined threshold can be highlighted. This can help to establish which genes are most changing over the time course. The data is scaled, so even low expression level genes (which many miRNAs are) can be contrasted without being marginalised.

One interesting contrast is the difference between the important genes identified in the source paper and the genes presented above [110]. These are different, because the source paper relied DE and ranking to identify interesting miRNAs, and while DE is a good metric, the *TimiRGeN R* package focuses on miRNA-mRNA interactions and time based changes, making it more specific.

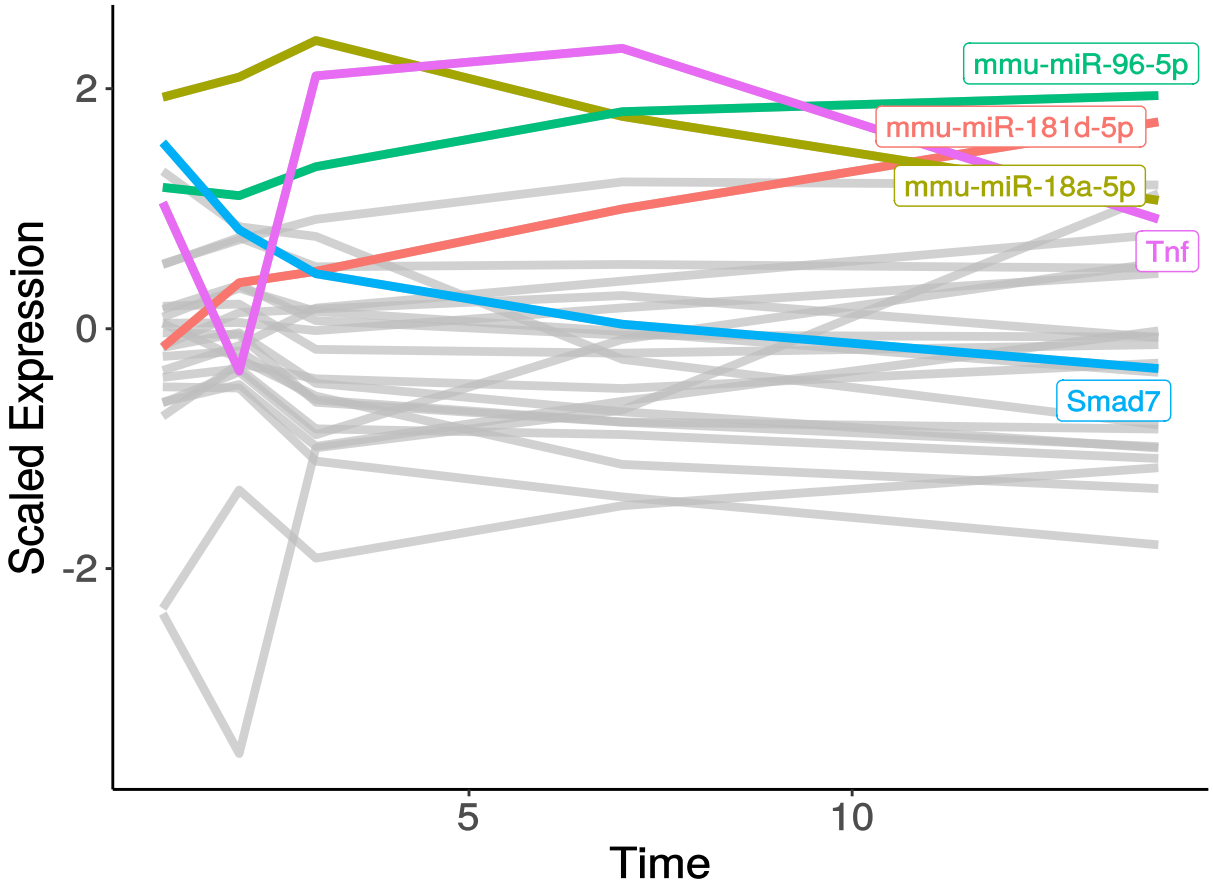


Figure 2.8: **Scaled gene behaviour over the kidney injury time course.** The time course shows several miRNAs and mRNAs which pass the threshold of 1.5 after scaling.

miRNA-mRNA pair analysis and cluster analysis methods

The genes within the network were also be explored using hierarchical clustering to identify any trends. No obvious trends or clusters came up. The genes involved in the twenty miRNA-mRNA interactions broadly fell into one of three clusters. These clusters are visualised by a heatmap or a dendrogram. These clusters can then be split into cluster specific line plots (Figure 2.9).

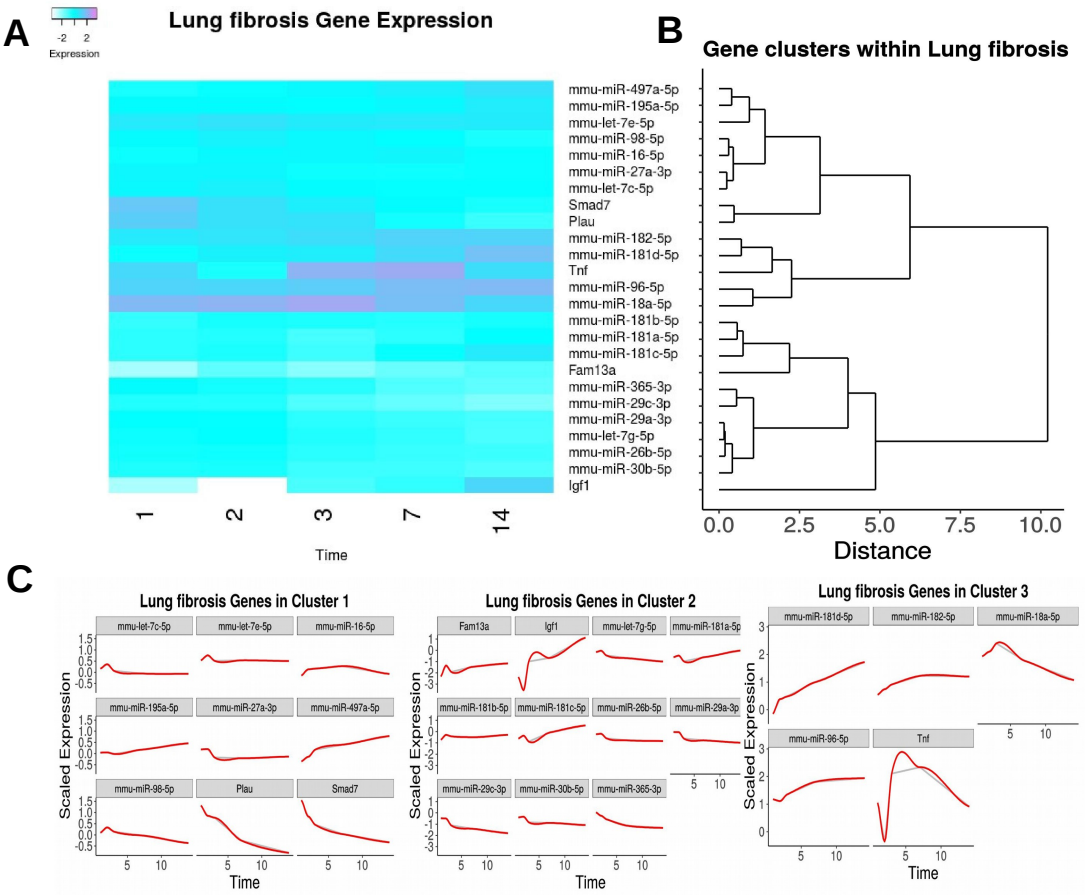


Figure 2.9: **Hierarchical clustering of miRNAs and mRNAs found after filtration.** Genes found to be involved in miRNA-mRNA interactions within the pathway of interest can be hierarchically clustered. **A)** A heatmap can be generated along with a **B)** companion dendrogram. Following this, a number of clusters can be defined by the user. **C)** Cluster based line plots can be displayed for each gene, per cluster.

Moving on, individual miRNA-mRNA pairs can also be explored and analysed using a suite of longitudinal pair analysis tools. To select a miRNA-mRNA pair to analyse, a heatmap (Figure 2.10) can be generated, and is ordered based on correlation scores between miRNA-mRNA pairs.

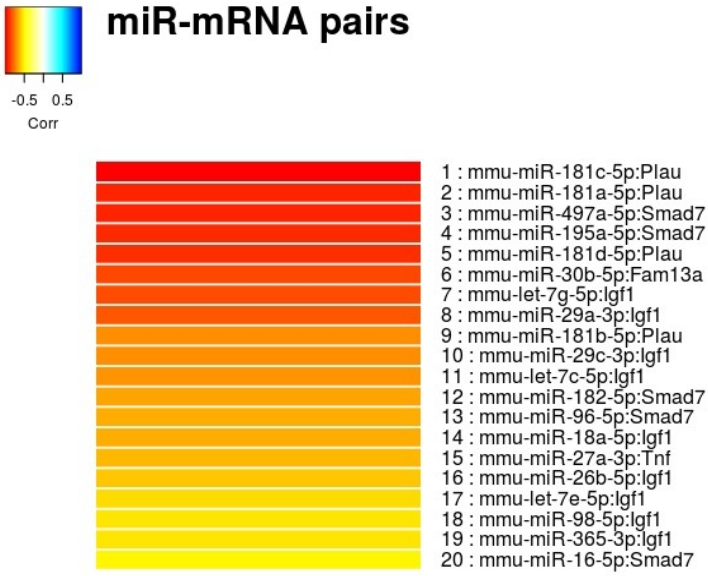


Figure 2.10: **Heatmap displaying miRNA-mRNA interacting pairs.** The heatmap is colour coded with a red-white-blue gradient. Red shaded interactions have a negative correlation, blue shaded interactions have a positive correlation and white shaded interactions have weak correlations. The heatmap indicates *mmu-miR-181c-5p-Plau* is the most negatively regulated interaction. All pairs have a correlation of at least -0.5, thus no white/blue shadings are seen.

The intention of this heatmap is to make selecting a miRNA-mRNA pair to further investigate simpler for users. As an example, the most negatively regulated interaction (*mmu-miR-181c-5p-Plau*) is used to display longitudinal miRNA-mRNA pair analysis methods from the *TimiRGeN R* packaged .

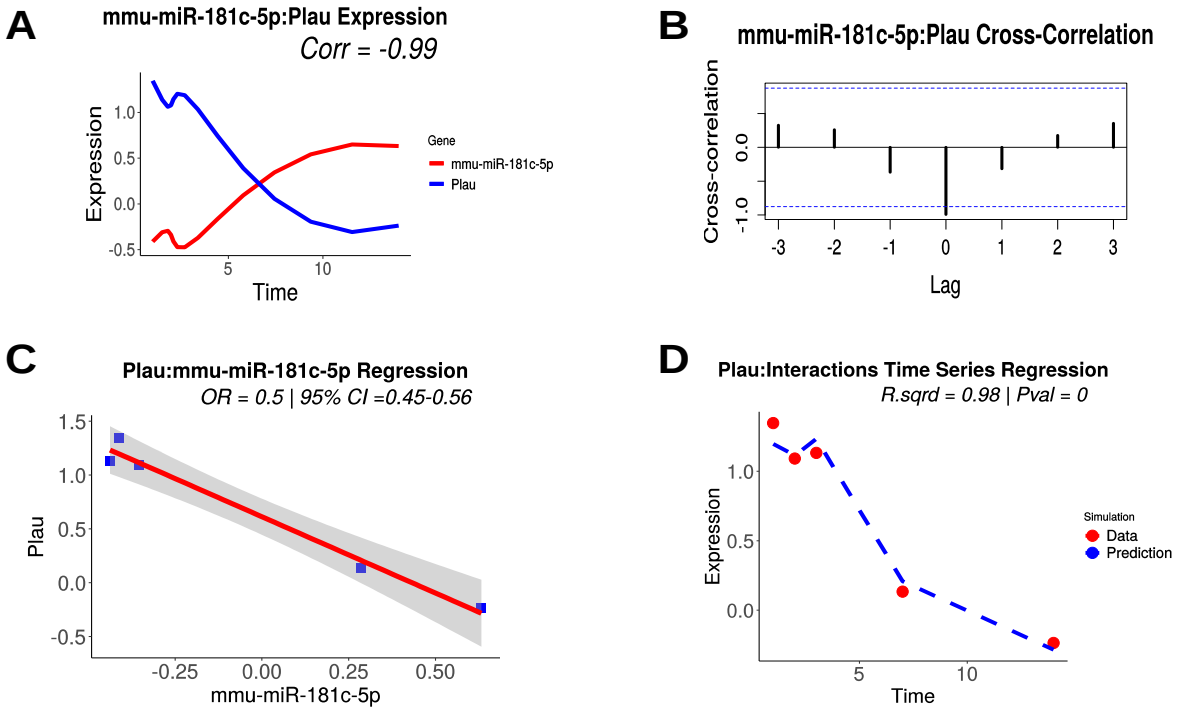


Figure 2.11: **miRNA-mRNA pair analysis metrics in the *TimiRGeN R* Package.** The most negatively correlated pair was selected, so this all plots here analyse *Plau* and *mmu-miR-181c-5p*. **A)** Line plot showing the miRNA and mRNA from a selected pair. This specific display is from interpolating the data over a smooth spline, and scaling log2FC values of the genes. The Pearson correlation is pasted as the sub-heading. **B)** A cross-correlation plot measuring the similarity between *Plau* and *mmu-miR-181c-5p*. It seems the two time courses are highly dissimilar due to the upwards and symmetrical sloping seen at the lags. This also indicates that the pair has an interesting dynamic. Interpolation was not used to make this plot. **C)** Simple regression between *Plau* and *mmu-miR-181c-5p*. OR and 95% CI are rounded to two decimal places. OR measures the likelihood of one time course effecting the other time course. An OR of 0.5 indicates there is a negative effect. CI gives a range (grey borders) where there is a 95% confidence that the mean of the data is within the range. An abline is also drawn. **D)** Regression plot showing predicted over measured data. Prediction by regression shows the time course of *Plau* levels predicted by *mmu-miR-181c-5p* levels (blue dashed line). This is in contrast to the measured *Plau* levels (red dots). R^2 and P value rounded to two decimal places are pasted as sub-headings.

2.2.5 GRN creation from Temporal clustering

The Lung Fibrosis pathway was found via time dependant pathway enrichment of the mouse kidney injury dataset (Figure 2.2A, Figure 2.3). *TimiRGeN* also offers a temporal clustering method to identify pathways of interest (Figure 2.2B). Temporal clustering is made possible by utilising the *Mfuzz* package [175].

This method compares the change in number of genes found to be significantly differentially expressed at each time point and each pathway. Soft clusters are used. Each pathway is given a membership score for each cluster (0-1), and the total score for each pathway will equal 1. Based on this score, fitness can be inferred, i.e. the higher a pathway scores with a cluster, the higher the fitness to the cluster. The colour of the lines reflects the degree of fitness a pathway has to a cluster; in order from the highest: red, orange, yellow, purple (Figure 2.12).

Cluster 1 was interesting because pathways with a high membership score to cluster 1 may be alternating in activity between days 3 and 7. Further investigation into pathways that fitted well into cluster 1 found the Inflammatory Response Pathway (IRP) (WP458).

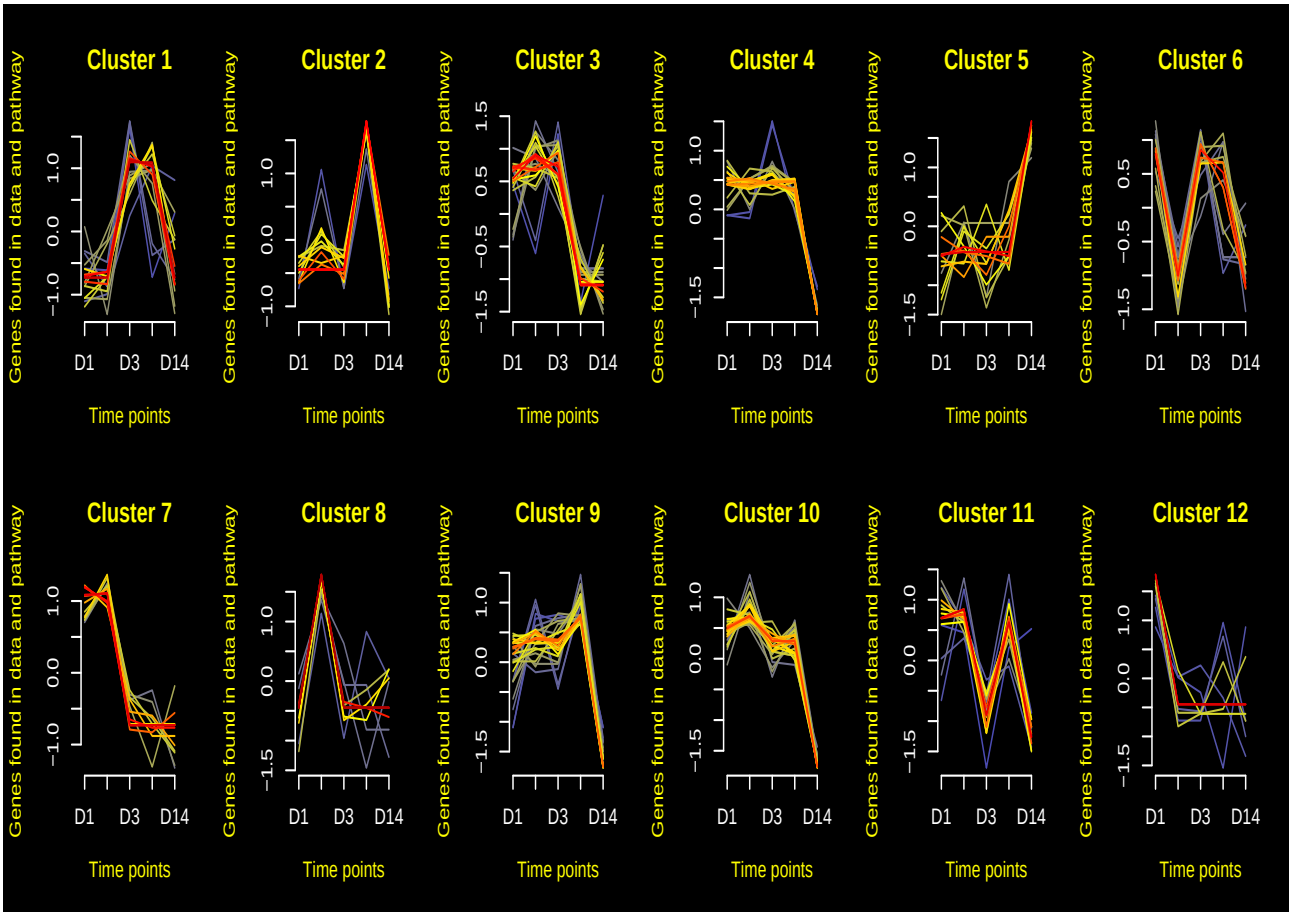


Figure 2.12: **Global analysis of miRNA-mRNA dataset using 12 clusters.** Fuzzy clusters have been used to cluster pathways based on temporal behaviours. The x-axis of each plot are time points and the y-axis are the standardized (between -1.0 and 1.0) number of genes shared by pathways and the input data. Pathways (lines) are colour coded, and from highest to lowest the colours are: red, orange, yellow and purple. The colours represent how well a pathway fits with a temporal behavior of a cluster.

Fourteen miRNA-mRNA interactions were found in the IRP after filtering for miRNA-mRNA interactions which were found in at least two miRNA target databases and had a Pearson correlation lower that -0.5. Most of the miRNA-mRNA interactions involved collagen promoting genes, which are known to be important factors in fibrosis progression, and fibrosis is a characteristic of kidney injury [176]. These miRNA-mRNA interactions were displayed in *PathVisio* to identify how the miRNAs regulate the IRP system. The resulting hypotheses were formalised into a GRN for collagen synthesis during kidney injury (Figure 2.16).

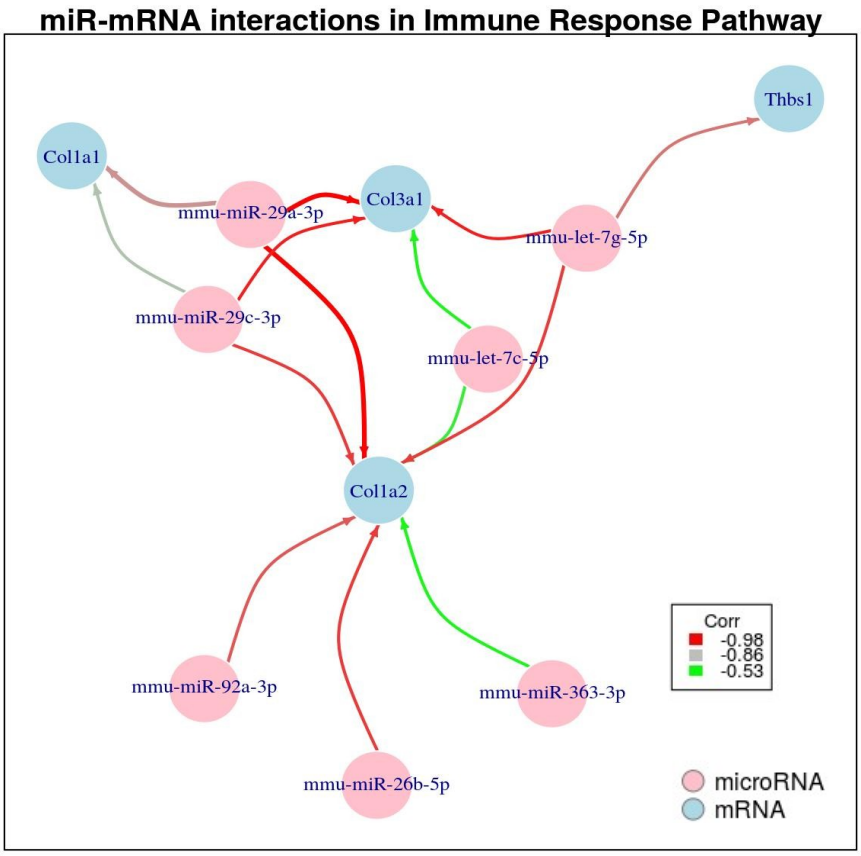


Figure 2.13: **Network showing miRNA-mRNA interactions found after filtering**. Made from all potential miRNA-mRNA interactions found in the Immune Response Pathway after filtration steps. All interactions in the plot have a Pearson correlation between -0.53 and -0.99. The colour of the edges represents the correlation and from highest to lowest the edges are coloured: green, grey, red. This image was edited to increase its visual quality.

The collagen synthesis GRN is induced by FA, which in-turn results in a fibrotic response. Here immune cell and structural repair gene activity leads to an increased rate of structural collagen production [177, 178]. Transcriptional activation of genes such as *Col1a1, Col1a2* and *Col3a1* contribute to structural extracellular fibers. This increases the rigidity of the surrounding tissue, leading to a loss of elasticity, which will negatively impact kidney function [176]. Each of the mentioned collagens increases over the 14 day time course and several of the predicted miRNAs decrease over the time course [109, 110].

The *TimiRGeN R* package predicted miR29 family miRNAs *miR-29a-3p* and *miR-29c-3p*

target each of the three structural collagen genes. Literature supports the miR29 family members regulating *Col1a1* and *Col3a1*, and on-top of this increasing miR29 family levels reduces fibrosis [179, 180, 181]. *TimiRGeN* predictions also identified *mmu-miR-29a-5p* and *mmu-miR-29c-5p* to regulate *Col1a2*, which has not been researched in renal systems. TimiRGeN also predicts Let7 family members regulated structural collagens after kidney injury. Specifically, *let-7g-5p* and *let-7c-5p* were predicted target *Col1a2* and *Col3a1* mRNAs. let-7d directly targets *Col3a1* [182]. Let7 members share most of their seed sequence, which further justifies the hypothesis of *mmu-let-7c-5p* and *mmu-let-7g-5p* targeting *Col3a1* mRNA [171].

Results also identify *mmu-miR-26b-5p*, *mmu-miR-92-3p* and *mmu-miR-363-3p* to target *Col1a2* mRNA. Only *miR-26b* had been experimentally validation of targetting *Col1a2* [183]. I believe that *mmu-miR-92a-3p* and *mmu-miR-363-3p* may be novel targets of *Col1a2*, under kidney injury conditions. This collagen synthesis GRN can be a resource for researchers investigating the regulation of structural collagens during kidney injury conditions. It can also lead to potential insights into non-coding RNA based therapies for CKD.

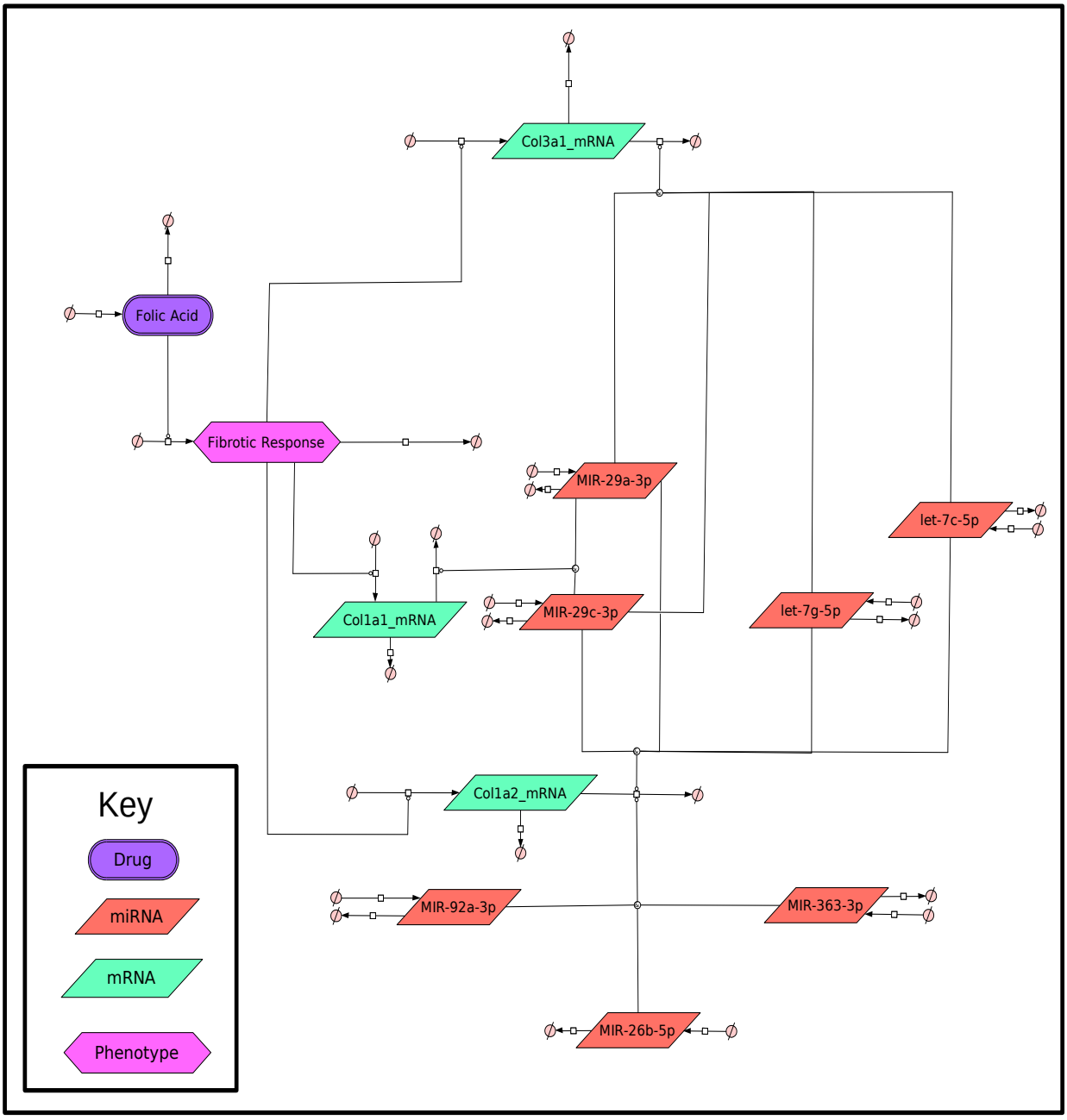


Figure 2.14: **GRN of miRNAs regulating collagen production.** miRNA-mRNA interactions found from the *TimiRGeN* package. Hypotheses about how the genes interact have been formalised into a GRN.

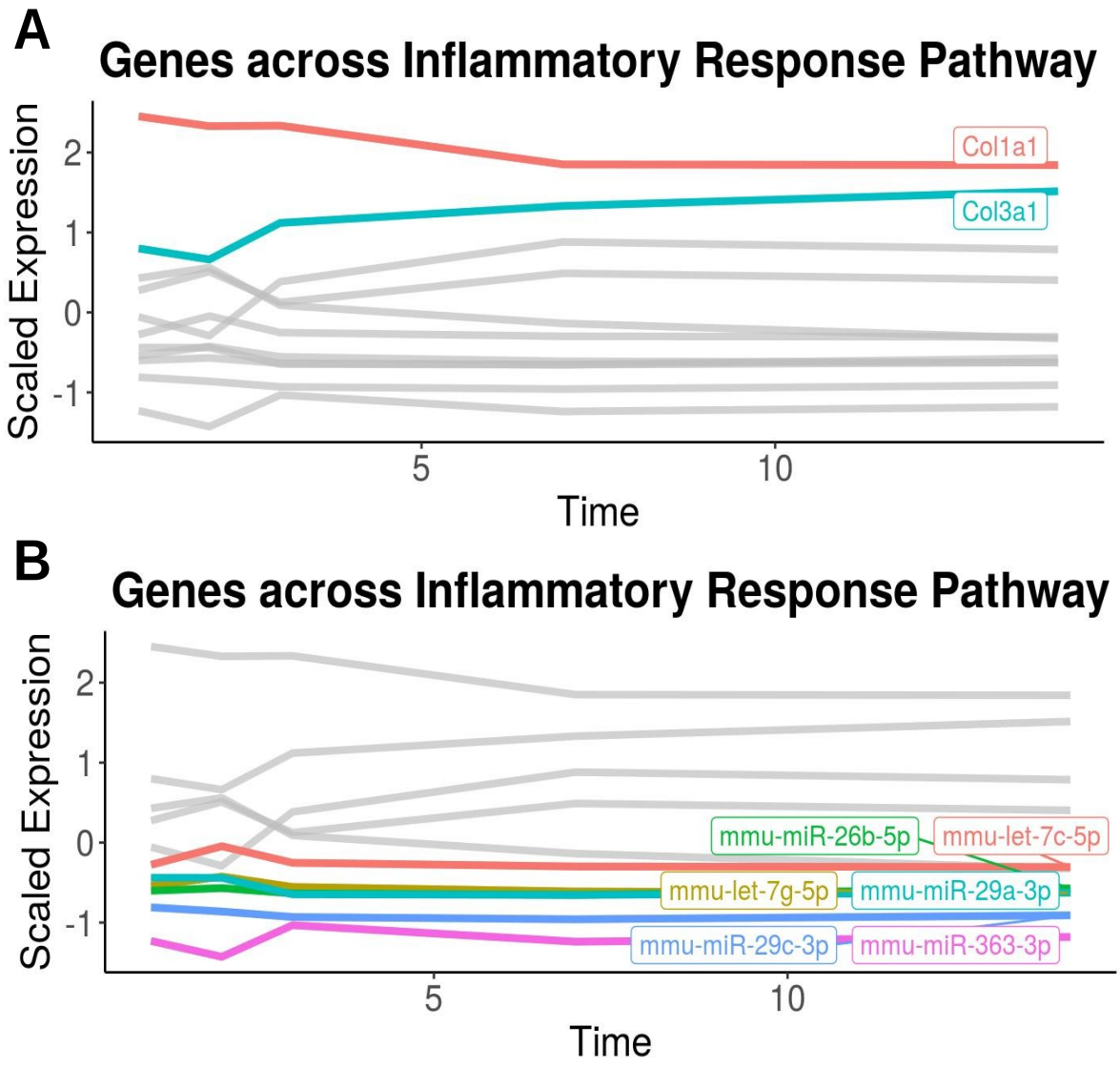


Figure 2.15: **Line plots showing the genes (miRNAs and mRNAs) found after filtration. A)** time series line plot which highlights genes which have a scaled value which surpasses 1 in at least on time point and **B)** time series line plot which highlights genes which have a scaled value which is lower than 0 in at least one time point.

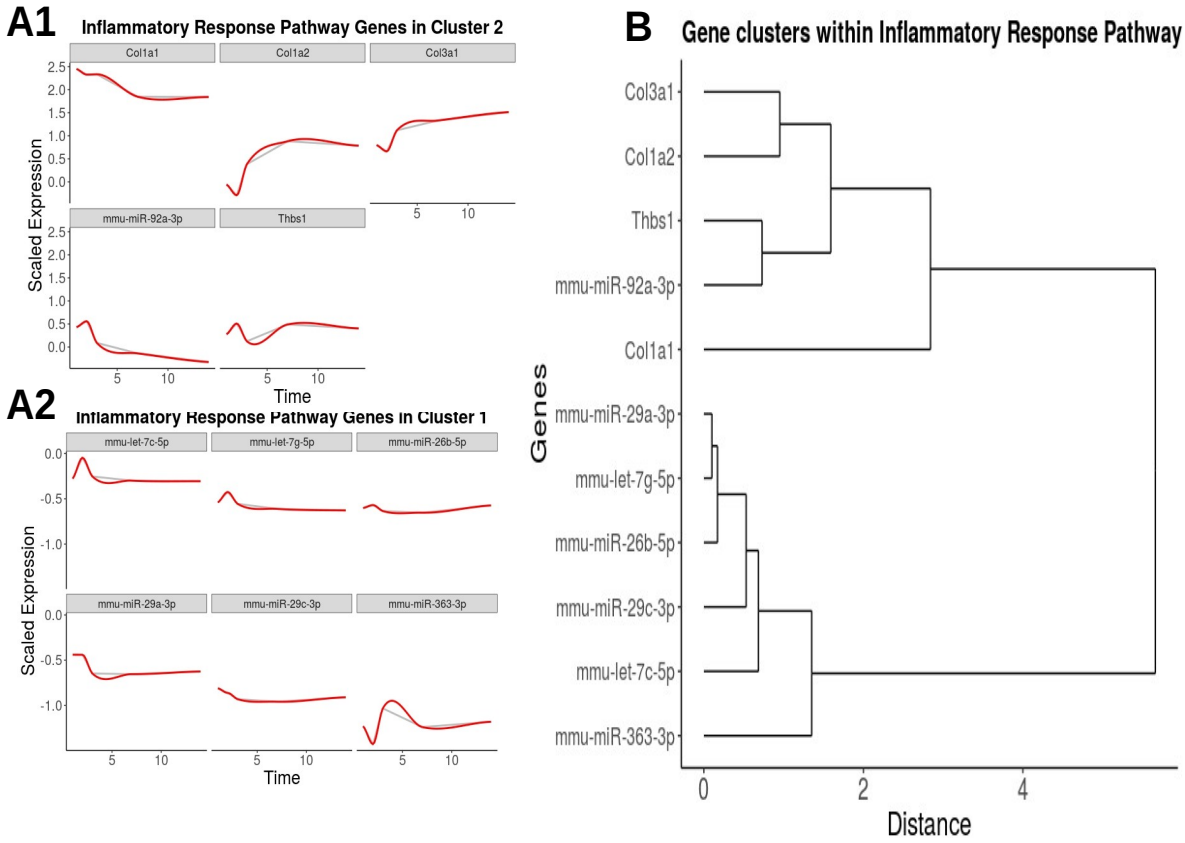


Figure 2.16: **Dendrogram and line plots from clusters 1 and 2 of the dendrogram.** Hierarchical clustering identifies two clusters, one which the collagen genes preside and one which contains most of the predicted miRNAs which target the collagens. **A1)** shows genes from cluster 2 and **A2)** shows genes from cluster 1. **B)** A dendrogram shows the distances of these genes.

Further exploration with *TimiRGeN* shows that *Col1a1* and *Col3a1* mRNAs are upregulated and several of the predicted miRNA partners are downregulated, possibly indicating that during the process of kidney fibrosis, the miRNAs are negatively regulated. Furthermore, hierarchical clustering shows the collagens and most of the predicted miRNAs form two separate clusters. A notable exception is *mmu-miR-92-3p* which is clustered with the collagens.

2.2.6 Alternate analysis methods with *TimiRGeN*

In subsection 2.3.4, I have discussed the use of this tool on a RNAseq based dataset which has had pairwise DE performed on it. Here I expand on the versatility of *TimiRGeN* by discussing how it can be used for: microarray datasets, non-pairwise DE input data and multivariate datasets. Again, all code linked to in Appendix C.

Analysis of microarray datasets

TimiRGeN works just as well with microarray datasets, as with RNAseq datasets. Users may wish to use a more specific set of genes as the background when performing over-representation analysis. For example, all known genes found within a cell or all probes in a microarray. The latter is an important factor when analysing microarray datasets for a more accurate analysis. As an example of this, a microarray hypoxia dataset was analysed with the *TimiRGeN R* package [112]. Probes were downloaded from platforms GPL6884 and GPL8227 and gene IDs extracted to create a list of genes for enrichment analysis. Separate miRNA-mRNA analysis mode was used, so each gene type (miRNA and mRNA) from each time point were analysed independently from each other gene type and time point. miRNAs are poorly annotated and often pathway enrichment with only miRNAs finds few pathways. The time points for the hypoxia dataset were: 0, 16, 32 and 48 hours after hypoxic conditions. Pairwise DE was performed using the 0 time point as the denominator. No enrichments were found at 16 hours. Results for 32 and 48 hours can be found in the github repository.

Averaged count/expression as input

All previous examples used pairwise DE because they are from shorter datasets. However for longer time series, alternative DE methods may be more appropriate such as using a cubic spline. There are many suitable time course DE tools e.g. *MaSigPro* or the *LRT* method in *DESeq2* [143, 145]. From these examples, a single log2FC and adjusted P value is given to each gene, so the standard input for *TimiRGeN* will not work. Instead, users will need to filter significantly differentially expressed genes from averaged count or

expression data, and use this as the input for *TimiRGeN* analysis. Generally, the functions are the same when using averaged count/ expression data from SDEGs or pairwise DE results (log2FC and adjusted P values) as the input, however the former input type has alternate options for selecting a pathway to investigate miRNA-mRNA interactions (Figure 2.17).

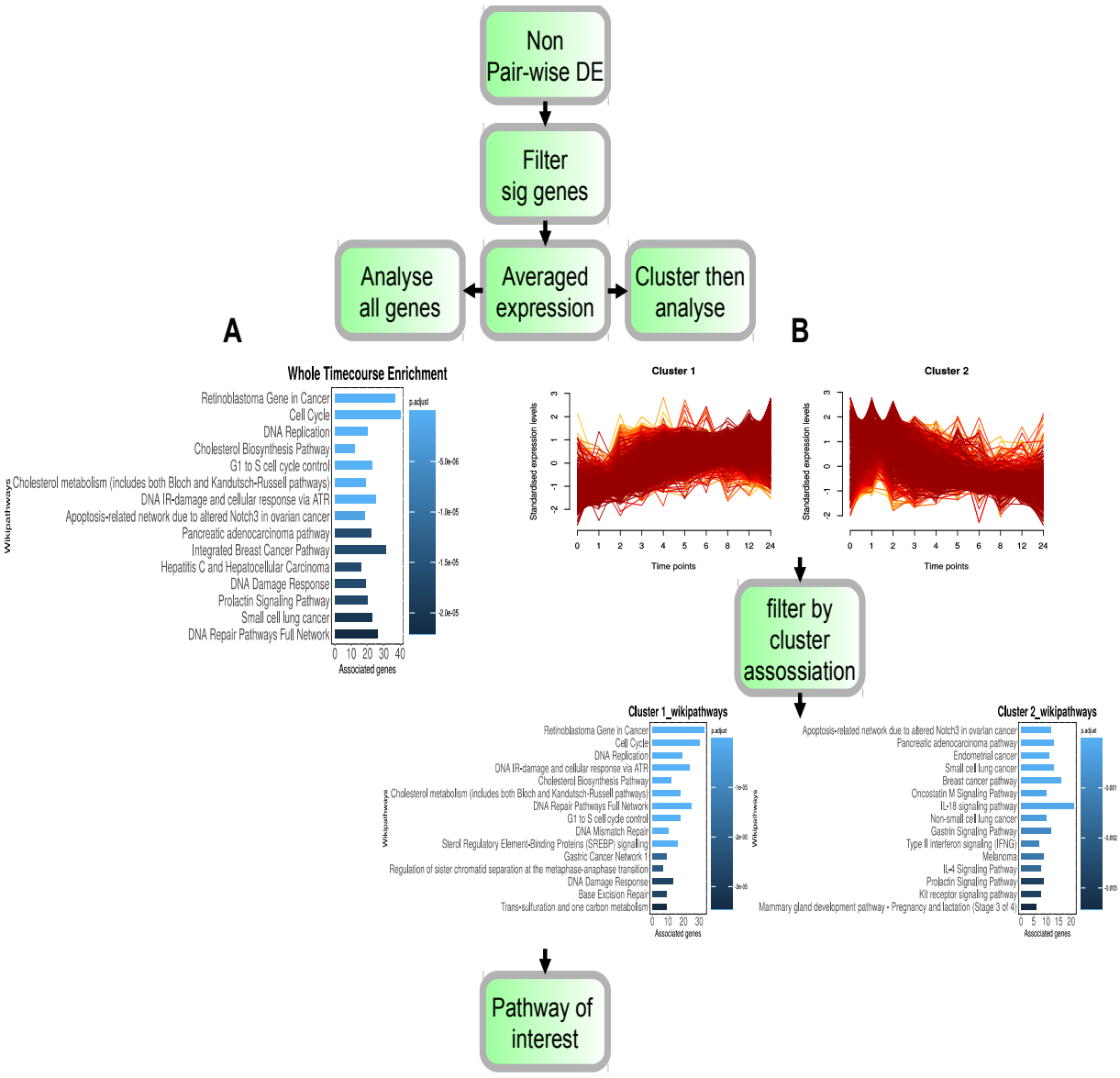


Figure 2.17: **Alternative *TimiRGeN* pipeline for non-pairwise DE.** Prior to using *TimiRGeN*, after a non-pairwise DE analysis is performed, the SDEGs will need to be filtered from averaged expression data. Following this, functional analysis can continue with *TimiRGeN* **A)** all genes can be functionally analysed with ORA to find pathways which are enriched over the whole time course or **B)** genes can be clustered first using fuzzy clustering, and then functionally analysed with ORA. After fuzzy clustering, genes can be filtered by their association with the clusters. The threshold of this association is user defined.

Multivariate datasets with multiple time series

More complex datasets will include multiple time series with multiple interventions. Such complexity can be explored with *TimiRGeN* by analysing each time series separately and then contrasting the miRNA-mRNA interactions found in each case. For this to work, the same pathway must be examined. As an example a second mouse kidney injury dataset was analysed. This dataset was generated from UUO, which is a surgical procedure involving the ureter connecting a kidney and the bladder being cut or blocked [184, 185]. This dataset was downloaded from GEO repositories GSE118340 (miRNA) and GSE118339 (mRNA). UUO data was processed the same way as the FA dataset. The Lung Fibrosis pathway was explored with both the FA and the UUO kidney injury longitudinal miRNA-mRNA datasets.

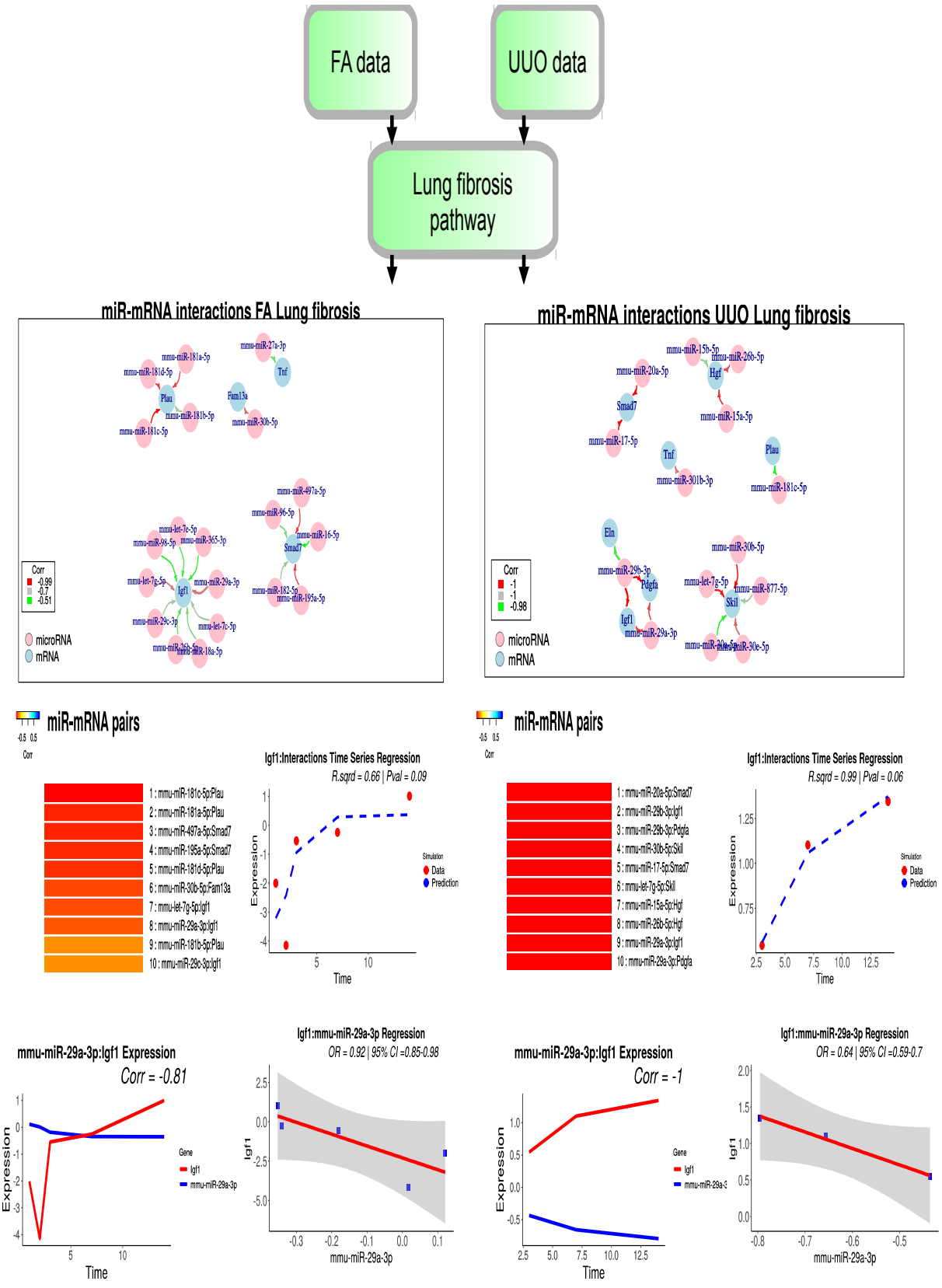


Figure 2.18: **Expanded *TimiRGeN* pipeline for multivariate datasets.** The FA and

UUO kidney injury datasets are contrasted using longitudinal pair analysis methods, within the context of the Lung Fibrosis pathway. Filtered miRNA-mRNA interactions from the FA dataset had a negative correlation of at least $<$ -0.5 and were found in at least two databases. Filtered miRNA-mRNA interactions from the UUO dataset had a negative correlation of at least $<$ -0.98 and were found in at least two databases. miRNA-mRNA interacting pair heatmaps show that the *mmu-miR-29a-3p-Igf1* interaction was found as the 8th and 9th most negatively regulated pair (according to Pearson correlation) in the FA and UUO datasets, respectively. Contrasting the longitudinal regression and correlation plots, shows the miRNA-mRNA interacting pair had a greater R^2, a lower Pearson correlation, and lower P value in the UUO dataset. Regression analysis shows the odds-ratio was higher in the FA dataset, also the CI range was larger in the FA dataset.

mmu-miR-29a-3p-Igf1 was found to be negatively correlated in both the FA and UUO datasets, so it is likely this interaction has a connection with kidney injury. It seems from the regression and correlation statistics that the pair has a greater role in UUO kidney injury, however there are only three time points in the UUO dataset. This may mean that the statistics are over-estimations, so therefore we can only predict *mmu-miR-29a-3p-Igf1* interaction may regulate the Lung Fibrosis pathway in both datasets, but cannot accurately assess whether the FA or UUO induced kidney injury models have the greater influence from the *mmu-miR-29a-3p-Igf1* interaction.

2.2.7 Publication

The *TimiRGeN R* package and associated results from the analysis were published in *Bioinformatics* in May 2021 as an open access original paper [117].

2.3 Methods

In this section I will describe in detail the necessary steps taken to produce a *Bioconductor* quality *R* package. I will also explain each of the functions within the *TimiRGeN R* package. Most of what is described in this section is from *R* and *Bioconductor* documentation [186, 187]. I also detail how the FA and UUO data were processed.

2.3.1 Data processing

FA was injected into the left kidney of mice. The specimens left kidneys were surgically removed and flash-frozen prior to, or 1, 2, 3, 7, 14 D after the FA injection. The kidneys had miRNA and mRNA measurements taken. No cell type was specified in the materials or supplementary materials so it is assumed a mix of multiple cells are used, and this was a tissue level study. The miRNA time course extended to D28, however the mRNA study did not, so the D28 data was not used. Data was downloaded with the *fastq-dump* function from *SRA-toolkit* [188]. Both miRNA and miRNA data were checked for quality using *FASTQC* [189]. The miRNA data had unneeded adapter sequencers which were removed with *cutadapt* [190]. For miRNA alignment the Mus_musculus.GRCm38.cdna.all.fa transcriptome was indexed by *Bowtie* and for mRNA alignment, *Salmon index* was used [191, 192, 193]. miRNA read quantification was performed with *miRDeep2* and mRNA quantification was performed with *Salmon* [194]. Reads were imported into *R* using *tximport* and pairwise DE was performed using *limma*. The zero time point was used as the common denominator during DE. UUO data was analysed in the same way as the FA data.

2.3.2 Package creation

TimiRGeN was created using *Roxygen2*, an *R* template generator which is recommended to use when creating new packages. *Roxygen2* automatically generates several files which are essential for an *R* package, and in this subsection I will briefly describe these different files, and some *Bioconductor* specifics. The package was created using *R* 4.0.2 and *BiocManager* 3.11.

The total amount of data contained withing a zipped package folder must be lower than 5 mb. Also, the total amount of time allowed for a package to build without vignettes must be less than 10 minutes. Exceeding these thresholds will results in *Bioconductor* warnings. This will also mean the package may need to be split into multiple smaller packages. At the time of writing the *TimiRGeN R* package, version 1.3.03 is 4.3 mb in size and took roughly 6 minutes to build the package without its vignette.

package tree

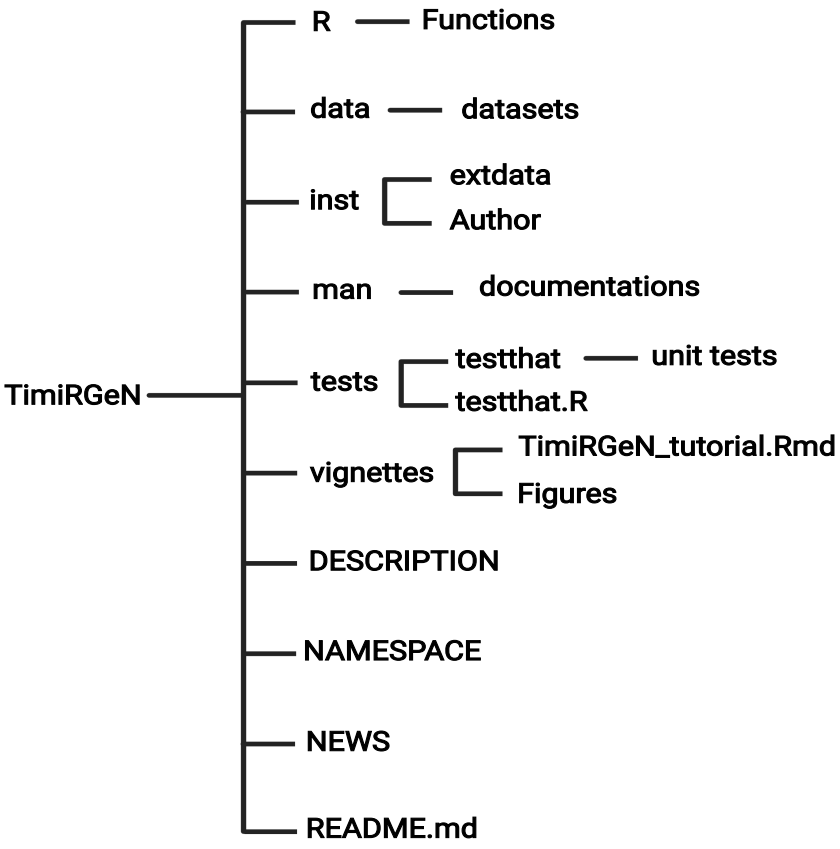


Figure 2.19: **Summary of the file structure needed for a *Bioconductor* tool.** This tree diagram depicts the order of folders and files which made the *TimiRGeN R* package.

One of the learning curves involved in creating a new *R/Bioconductor* package is file structure. It requires many specific features to pass builds and checks. Figure 2.19 shows

the structure of the package accepted by *Bioconductor*. The folders and individual .R, .rda or .Rd files. Each part will be explained below.

data

The data folder includes ten .rda files which the end user may want to use. Including six example longitudinal miRNA-mRNA datasets which help by providing templates of how the input data should look like. The datasets comprise of samples from the: 1) FA fibrosis dataset as the main example file (miRNA), 2) FA fibrosis dataset (mRNA), 3) the UUO fibrosis dataset to show users how to perform multivariate analysis (miRNA and mRNA data were kept together), 4) the hypoxic breast cancer dataset to show how a human dataset can be analysed and as an example for microarray based datasets (miRNA, 5) hypoxia dataset (mRNA), 6) another breast cancer dataset used as an example for non-pairwise DE samples (miRNA and mRNA data were kept together) [109, 110, 111, 112, 184]. These example datasets were a subset of the original datasets, and this helps examples run faster and speed up the vignette building process. Reducing dataset size also helped to make the tool smaller.

The other four .rda files help run functions which requires downloading information, which does not always work during *Bioconductor* vignette building. To simplify this, much of the data that would be downloaded during the vignette building process are stored in the data folder instead. Datasets which are not essential for the end user to see (not used in the examples) are stores in the /Ints/extdata folder. A complementary Rd. file is stored in the man folder for each of the ten dataset here. Users can read descriptions of the man files to find which dataset are most appropriate to help their analysis.

DESCRIPTION

The DESCRIPTION file holds some of the most important information in an *R* package. I have included the code for DESCRIPTION file of *TimiRGeN* v1.3.03 below, and will describe the relevance of each section.

```
‘‘‘

Package: TimiRGeN
Type: Package
Title: Time sensitive microRNA-mRNA integration, analysis and
          network generation tool
Version: 1.3.03
Authors@R: person(given = "Krutik", family = "Patel",
                  role = c("aut", "cre"),
                  email = "K.Patel5@newcastle.ac.uk")
Description: TimiRGeN (Time Incorporated miR-mRNA Generation of
             Networks) is a novel R package which functionally
             analyses and integrates time course miRNA-mRNA
             differential expression data. This tool can generate
             small networks within R or export results into
             cytoscape or pathvisio for more detailed network
             construction and hypothesis generation. This tool
             is created for researchers that wish to dive deep
             into time series multi-omic datasets. TimiRGeN
             goes further than many other tools in terms of data
             reduction. Here, potentially hundreds-of-thousands
             of potential miRNA-mRNA interactions can be whittled
             down into a handful of high confidence miRNA-mRNA
             interactions affecting a signalling pathway, across
             a time course.
License: GPL-3
Encoding: UTF-8
LazyData: true
RoxygenNote: 7.1.1
Depends: R (>= 4.0.2),
         Mfuzz,
         MultiAssayExperiment
Imports: biomaRt,
```

```
        clusterProfiler,
        dplyr (>= 0.8.4),
        FreqProf,
        gtools (>= 3.8.1),
        gplots,
        ggdendro,
        gghighlight,
        ggplot2,
        graphics,
        grDevices,
        igraph (>= 1.2.4.2),
        RCy3,
        readxl,
        reshape2,
        rWikiPathways,
        scales,
        stats,
        tidyr (>= 1.0.2),
        stringr (>= 1.4.0)
Suggests:
        BiocManager,
        kableExtra,
        knitr (>= 1.27),
        org.Hs.eg.db,
        org.Mm.eg.db,
        testthat,
        rmarkdown
VignetteBuilder:
    knitr
biocViews:
    Clustering,
    miRNA,
    Network,
```

```
    Pathways,
    Software,
    TimeCourse,
    Visualization
URL: https://github.com/Krutik6/TimiRGeN/
BugReports: https://github.com/Krutik6/TimiRGeN/issues
```

The Package, Type of package, Title, Authors and Description sections are self explanatory. The Version is important for updating the package in *Bioconductor*. *Bioconductor* updates must follow special bump X.Y.Z rules. Currently, the package is version 1.3.03, for a successful version bump the Z must be increased by an increment of one, so 1.3.04. Anything else, would not work. License being GPL-3 means that *TimiRGeN* is a free to download and modify. Encoding is the character coding type used, which is the standard UTF-8. LazyData being true means that datasets will be downloaded only after they are called, so datasets attached to the package will not take up memory when they are not in use. RoxygenNote is the version of *Roxygen* used in development. Other sections of note are the VigneteBuilder which indicates the method of vignette building and biocViews, which is a list of *Bioconductor* key words. This list will categorize this package to help potential users find it.

The DESCRIPTION file also allocates other packages needed for *TimiRGeN* to function. These packages will be divided between Depends, Imports and Suggests. Packages and software in the Depends section are essential for *TimiRGeN* to function and are downloaded and loaded onto the search path when *TimiRGeN* is loaded, i.e. there is no need to call for their libraries. Here the *R* version is present, so users will need *R* 4.0.2 (or newer) to install *TimiRGeN*. *TimiRGeN* is also dependant on *Mfuzz* because one of the *Mfuzz* dependencies, *e1071* must be on the search path for *Mfuzz* functions to work [152, 195]. The simplest solution to get *e1071* on the search path was to place *Mfuzz* in the Depends section, and thus *e1071* is indirectly kept in the search path of *TimiRGeN*. *MultiAssayExperiment* is also in this section because most functions use this package, so

it would not make sense to need to call for *MultiAssayExperiment* when using *TimiRGeN* [196]. Packages in the Imports area will also be downloaded when *TimiRGeN* is downloaded but not automatically be put on the search path. This is because these packages are needed for some of the function in *TimiRGeN*, but not all. So these packages are only called when the functions required to use them are called. When functions which use packages in the imports list are called, either the attached package is loaded, or specific functions from the attached package are loaded, depending on the way the specific functions are coded. This makes the *TimiRGeN* package smaller and take less space. This also makes some functions work faster, as they may only require a specific function from a package on the Imports list. Finally, packages in the Suggests section are not loaded when when *TimiRGeN* is called. Packages here are not essential for *TimiRGeN* to function e.g. if a user is interested in analysing *Homo sapiens* data, they would not need to download *org.Mm.eg.db*. This would be a waste of time and space. Users will have to download and call for packages in the suggests list themselves.

inst

The inst folder contains extra material which may help the package but the contents are not essential for the package to run. It holds an extdata folder. In here 7 .rda files are found. These are not essential for the end user to see. These only serve to speed up the vignette building process. Having these datasets here, rather than in the data folder, means no .Rd files (contents of the man folder) need to created, and so the package building process is faster and less memory is used. Finally, an author file and a CITATION file are found here. The CITATION file is written in a bib code style.

man

The man file contains documentation of the functions from the R folder and the datasets from the data folder. When *TimiRGeN* is loaded, a user can use the help panel to gain more knowledge about certain functions or datasets. The help pages will be from the .Rd files stored in the man folder. The .Rd files are automatically generated by *Roxygen2* from the functions and datasets. Functions with a @noRd annotation will not generate an .Rd

file.

NAMESPACE

An automatically generated *Roxygen2* file containing all the *TimiRGeN* functions, functions required from attached packages, attached packages and datasets in the data folder, and all of these are exported with *TimiRGeN*. At the time of writing, a total of 58 symbols are exported - including 10 datasets and 48 functions; making *TimiRGeN* a large package.

NEWS

This is a regularly updated document which states the data and version change, and bullet points the changes associated with each new version.

R

The R folder contains all the functions in the *TimiRGeN* package as .R files. A total of 55 functions are in the TimiRGeN package, 7 of which are internal functions and not exported in the NAMESPACE file and .Rd files are not generated for these 7 functions. Each function follows a similar Roxygen skeleton.

```
#' @title
#' @description
#' @param
#' @return
#' @export
#' @importFrom
#' @import
#' @usage
#' @examples
Function
```

Functions not exported to the end user will have the @noRd attribute instead of the @export attribute.

```
#' @noRd
```

The @title and @description are self explanatory. @param will describe the function parameter, and every parameter has to be defined with a new @param section. @return will describe the output of the function and @export will add the function to the NAMESPACE. @importFrom states if functions from another package are required for the function to work. Imported functions from other packaged are called as so "package::function". Some functions use @import to load the whole package for use in a function. @usage describes how the function can be used, so when describing @usage, all parameters must be present in the same order they are called for in the function. @examples shows a working example of the function. Some demanding functions have their @examples sped up by importing a dataset from the data folder. Some functions have @examples blank because they are too time consuming. In a *Bioconductor* package, at least 80% of functions need to have working examples presented. The Function is presented in full at the end of the .R file.

README.md

A markdown file which gives a short description about the package, and can be seen as the introduction to the package. This will function as the main text to go on the front of the package's associated github website (https://github.com/Krutik6/TimiRGeN/).

tests

This folder contains an .R file which calls for the *testthat* package to run all the unit tests found in the tests/testthat folder [197]. The testthat folder contains *R* scripts and .rda files to unit test functions. A total of 127 tests are performed on this package to make sure

most of the functions are working. Not every function is tested this way due to size and time constraints, but many of the central functions are tested.

vignettes

This folder hold the TimiRGeN_tutorial.Rmd file which is the package vignette. This holds a number of working examples and in depth explanations on many of the nuances of the package which would be difficult to understand without the working examples. Alongside the .Rmd file are 15 png files which help showcase the output of the package.

2.3.3 Functions

I will list all the functions of the *TimiRGeN R* package in alphabetical order, and briefly describe them, the intended output and the rationale of each functions. The full code for the functions can be found in the R folder in the github repository https://github.com/Krutik6/TimiRGeN/R, which is linked to in Appendix C. Some functions will specify if there are differences between combined (c mode) or separated (s mode) of miRNA-mRNA data analysis. Other functions will only be needed for non-pairwise DE. Internal functions are also included because internal functions help other functions work.

addIds

addIds maps retrieved entrezgene or ensembl gene IDs from a getIds function to dataframes containing significantly differentially expressed genes. *addIds* will work on a nested dataframes (c mode of analysis) or on a list of nested dataframes (s mode of analysis).

addPrexif

addPrexif is a necessary function for s mode of analysis. This function will add a user defined prefix to each column name of the input files. With this, downstream functions will be able to differentiate between the miRNA and mRNA data, and analyse them separately.

clusterPrep

clusterPrep is an internal function which prepares data for hierarchical clustering. This

function uses *reshape2::melt* to prepare data for plotting [198].

clusterCheck

clusterCheck creates a PCA plot by wrapping around the *Mfuzz::overlap* function [152]. This will compare the distances between cluster created by the *createClusters* function. With this, users can see if they have created too many, or too few clusters. This function uses the *grDevices::dev.new()* function to give the option of plotting in a new window [199].

clusterList

clusterList is a function designed to be used when analysing longitudinal miRNA-mRNA datasets analysed with non-pairwise DE approaches. *clusterList* will transform clusters created by the *createClusters2* function into lists based on genes associated to each cluster. Genes association with clusters are determined by the *fitCluster* parameter in the function. Users can define their own threshold (0-1). The default *fitCluster* threshold is 0.5.

combineGenes

combineGenes is a necessary function for the c mode of analysis. The miRNA and mRNA data will be combined into one dataframe. Both datasets must have the same column names. To make sure a chronological order is maintained in the column names, *gtools::mixedsort()* is used [200].

corrTable

corrTable is an internal function which helps the *mirMrnaInt* function create a miRNA-mRNA correlation matrix. It uses the *R stats* package to generate correlations between each miRNA and mRNA inputed into the *mirMrnaInt* function. Pearson correlation is the default method but Kendall and Spearman are available options.

createClusters

createClusters uses time course data to generate temporal fuzzy clusters. If s analysis is performed, the prefixes defined in the *addPrexif* function can be used to analyse the sig-

nificantly differentially expressed miRNA or mRNA data; one at a time and individually. Or if c analysis is performed, all the significant differentially expressed genes are used at the same time. The input of this function is a percentage matrix which indicates the percentage of genes in each WikiPathway at each time point, and this percent matrix is generated by the *percentMatrix* function. The following functions are used: *Mfuzz::filter.std , standardise, mestimate, mfuzz* to convert the percentages into standardised values between 0 and 1 [152]. The changes in the number of genes in each WikiPathway over the time course are used to determine different temporal patterns. These patterns are divided into *k* clusters. *k* is selected by the user.

A priori filtration step to reduce the number of pathways which do not show significant variance in the number of genes across the time course is taken. The extent of the removal can be controlled by the user, and this can be visualised with a standard deviation plot. Remaining pathways have their details downloaded by using *rWikiPathways::getPathwayInfo* [124].

As described in *Futschik et al (2005)*, there are some important factors involved in calculating soft clusters [201]. Firstly, the overall equation for c-means soft clustering is the equation below.

$$Msc = \left\{ \begin{array}{lll} Uij \quad \epsilon \quad R^{N.c} & | & Uij \quad \epsilon[0,1] \forall i,j \\ & | & \sum^{c} j = 1 \quad Uij = 1 \forall i \\ & | & 0 < \sum^{N} i = 1 \quad Uij < N \forall j \end{array} \right\}$$

Here each ith term is a pathway and each jth term is a cluster. Uij is the membership score of a pathway has for a cluster. N denotes the number of objects within the analysis, c is the number of soft clusters and $\epsilon[0,1]$ forces Uij to be equal a value between 1 and 0. This information helps to identify the soft partitions between clusters (Msc).

Several parameters are calculated to support c-means fuzzy clustering. Jm is calculated

to weigh distance of data (pathways) xi to the center of a cluster cj. This function takes into account the membership values of the data point (xi).

$$Jm = \sum_i \sum_j (Uij)^m ||xi - cj||_A^2$$

m is a parameter calculated by a quadratic equation between distances of data objects. A is also related to this quadratic equation [201]. Jm is used to determine m, which holds power in determining the influence of data points (xi) during clustering (calculating Msc). If the data is very noisy, m will be a large value, and this will lead to poorly clustered (potential outliers/ noise) to having small Uij values, and thus the data point (pathway) is having a smaller affect on cluster partitioning (Msc). This makes Jm calculation a necessary noise-reduction step. Filtration of pathways/ genes with low variance will likely positively impact the objective function of Jm.

createClusters2

createClusters2 will create temporal clusters for input data which has not come from pairwise DE. This input data should be averaged count/ expression data across a time course. During plotting, unlike with *createClusters*, each line will represent a gene, rather than a pathway. In contrast with *createClusters*, each ith term is a gene and Uij is the membership score a gene has for a cluster.

cytoMake

cytoMake exports filtered miRNA-mRNA interactions from *R*, into *Cytoscape*. For this to work, *Cytoscape* version 3.7 or later must be open, and *RCy3::cytoscapePing()* must be used first [154]. In this function, *RCy3::createNetworkFromDataFrames, setVisualStyle, layoutNetwork* are used.

dataMiningMatrix

dataMiningMatrix adds databases as columns to the correlation matrix made by the *mirMranInt* function. In these columns, 1's represent miRNA-mRNA interactions being found

in a database, and 0's represent interactions not being found in a database. Score summaries are also added. This function can work with up to three databases (TargetScan, miRDB and miRTarBase), and can work even if some databases are not successfully downloaded.

diffExpressRes

diffExpressRes extracts annotation IDs (ensembl or entrez) and a single type of DE result which is uniformly available for each sample, preferably one which represents magnitude of change e.g. log2FC. This function must be performed for miRNA and mRNA data separately.

dloadGmt

dloadGmt downloads species specific wikipathway information, including: entrezgene IDs, pathway descriptions and pathway IDs. This function uses: *clusterProfiler::read.gmt*, *rWikiPathways::downloadPathwayArchive*, *tidyr::separate* and *dplyr::select* [124, 202, 203, 204]. If the most recent species specific gmt file is not downloadable, a recent (within 8 months) version of the file will be downloaded instead. At the time of writing, the march 2021 versions are used as backups.

dloadMirdb

dloadMirdb downloads most recent version of miRDB data (version 6.0), for a specific species [79]. The downloaded file will be formatted to be usable by *dataMiningMatrix*.

dloadMirtarbase

dloadMirtarbase loads the most recent version of miRTarBase data (version 8.0), for a specific species [72]. Formatted *dloadMirtarbase* data is already loaded within *TimiRGeN* because its download server is very slow and unstable. miRNA-mRNA interactions which are labelled as "weak" evidence have been removed.

dloadTargetscan

dloadTargetscan downloads the most recent version of TargetScan data (version 7.2), for

a specific species [83]. The downloaded file will be formatted to be usable by *dataMiningMatrix*.

eNames

If using c mode of analysis *eNames* will generate a list of significantly differentially expressed gene IDs found at each time point. Or if using s mode of analysis, *eNames* will generate a list separated by gene type (miRNA or mRNA), and these will subsequently contain lists of significantly differentially expressed gene IDs found at each time point.

enrichWiki

enrichWiki uses an ORA method. This uses hyper-geometric tests to identify enriched WikiPathways for each time point. This method uses the *enricher* function from *clusterProfiler* [202]. For s analysis, this function will find enriched pathways for the miRNA and mRNAs separately. The equation below has been taken from a website created by the author of *clusterProfiler*, Guangchuang Yu [205].

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Putting this equation into context of *TimiRGeN*, this function relies on a universe of unique background genes *N*. The default *N* are unique gene IDs found in all WikiPathways. With this in mind, the default question is which WikiPathways are most enriched based on the input data? A user could also define their own *N*. This is useful for a more stringent test, or if the input data is from a microarray dataset, the user may want to change the universe to consist of only genes probed for on the specific microarray platform. *M* represents the genes within *N* that are annotated in an individual WikiPathway. The number of SDEGs within a given time point is *n*, and *k* represents genes found within *n*, that are annotated in a specific WikiPathway (M). Each WikiPathway will be analysed individually. Following this, plots can be generated using the *quickBar* function to display genes with the lowest adjusted P values based on ORA.

genesList

genesList creates nested lists out of dataframes. Either nested lists based on time points (c analysis) or nested lists based on gene type which are then further nested based on time points (s analysis). This function used *gtools::mixedsort* to organise data and *stringr::str_extract* to separate the prefixes during s analysis [206, 207].

getIdsMir

getIdsMir will retrieve entrezgene IDs and ensembl gene IDs for miRNA data. It also produces adjusted gene IDs for both ID types. This can help to distinguish miRNAs that share a common ID when exporting data or when generating networks. *clusterProfiler::bitr* is used and the function currently works on many vertebrate model organisms including human, mouse, rat and zebra fish [202].

getIdsMrna

getIdsMrna will retrieve entrezgene IDs and ensembl gene IDs for mRNA data. Here, *biomaRt::useEnsembl* and *biomaRt::getBM* are attempted in the first instance [208]. If connection to *biomaRt* is unavailable, then *clusterProfiler::bitr* is used instead [202]. Generally, *biomaRt* will lead to a greater number of annotations for mRNAs, but the connection to the server can often fail due to server issues. This function also allows users to attempt different *biomaRt* server mirrors. Connection success/ failure to *biomaRt* is reported to users.

getP

getP calculates P values from linear models generated by the *quickTCPred* function. This is an internal function. The P value will be pasted onto plots to show significance of miRNA-mRNA interactions after predictive regression analysis. The f-distribution test is calculated in *R* by the *stats:pf* function and it uses the following parameters from a linear regression model: value, degree of freedoms of the numerator, degree of freedoms for the denominator to calculate the p-value.

$$H_o = u1 = u2$$

The null hypothesis of the f-distribution test is that there is no different between the samples (mRNA ($u1$) and miRNA ($u2$)). The f-distribution test calculates the likelihood of the f-statistic found from linear regression to be correct.

gmtEnsembl

gmtEnsembl converts the entrezgene IDs from the *dloatGmt* function into ensembl gene IDs by using *clusterProfiler::bitr* [202]. Most *WikiPathways* are annotated with either entrezgene IDs or ensembl gene IDs. With this, both are available for downstream analysis. If a selected Wikipathway is imported into *PathVisio* for further analysis, a compatible annotation type is needed.

hClustPrep

hClustPrep is an internal function which further helps to separate data into hierarchical clusters. This function is used in synergy with the *clusterPrep* function. ”maximum” is the default distancing method and ”ward.D” is the default hierarchical clustering method.

linearRegr

linearRegr is to be used after the *multiReg* function. *linearRegr* will generate linear models between the selected gene of interest, and any number/ combination of the genes predicted binding partners. Unlike most other functions, *linearRegr* will not generate a MAE object because users may wish to create and test multiple models.

makeDynamic

makeDynamic generates a file which contains information to import dynamic visuals into *PathVisio*. This includes gene names (miRNA and mRNA) and dynamic information in a chronological order. The dynamic information is a result type from DE; the most appropriate is log2FC, but normalised expression can also be used. This file also includes gene IDs. Either entrezgene or ensembl. It is recommended to also include the adjusted miRNA IDs in case the predicted miRNA-mRNA interactions contain miRNAs with shared genomic IDs. Finally, a system code is included for each gene entry. This is a specific ID label used in *PathVisio* [124]. This will be ’L’ for entrezgene IDs or ’En’ for ensembl gene

IDs. Users check the ID types used to annotate the selected pathway of interest. Missing annotation IDs may have to be manually entered.

makeMapp

makeMapp generates a file which can be imported into *PathVisio* using the Mapp builder app [124]. Doing so will import all the filtered miRNAs, associated adjusted IDs and a system code ('L' or 'En'). Once imported, the user will have to manually match the miRNAs and their targets on *PathVisio*. Missing annotation IDs may have to be manually entered.

makeNet

makeNet converts filtered miRNA-mRNA interactions into an igraph format using *igraph::graph_from_data_frame*; from which an internal *R* network can be visualised [153].

matrixFilter

matrixFilter filters miRNA-mRNA interactions from a large correlation matrix. The filtration options for users are :

- Negative correlations only?
- Maximum correlation allowed? (-1 to 1)
- Predictive databases only? (TargetScan and miRDB only)
- Minimum databases which an interaction has been found in? (0-3)

micrornaFull

micrornaFull is an internal function which is used by *getIdsMir* to standardise miRNA names for compatibility with *clusterProfiler* [202].

mirMrnaInt

mirMrnaInt uses the *corrTable* internal function to create a correlation matrix between miRNAs and mRNAs (found in common between input data and pathway of interest). Correlations are created by taking into account, chronological changes in magnitude over

time. log2fc is an appropriate input for correlations, but averaged count/ expression can also be used, especially if users use non-pairwise DE methods.

multiReg

multiReg extracts and formats data for the *linearRegr* function. A selected gene (mRNA or miRNA) and its predicted binding partners are exported into a new matrix.

nonUnique

nonUnique in an internal function which uses the *R stats* package (*stats::ave*) to create the adjusted IDs for the *getIdsMir* function.

pickPair

pickPair is an internal function used in numerous plotting functions where a single miRNA-mRNA pair needs to be selected for analysis.

quickBar

quickBar generates a bar plot from the outcome of ORA via the *enrichWiki* function. This function uses *ggplot2::ggplot* [209, 210]. The x axis will show number of genes, and y axis shows WikiPathways. The plots are colour coded to represent adjusted Pvalues. Size of plots and number of outputted pathways can also be altered by the user.

quickCrossCorr

quickCrossCorr will perform cross-correlation analysis on two time series (a miRNA and a mRNA). Cross-correlation uses *stats:ccr* and *scales::rescale*. Interpolation and scaling are possible. *quickMap* is recommended to be used first as this will order miRNA-mRNA pairs by descending correlation, and thus make selection of pairs easier.

Cross-correlation is a normalised version of the cross-covariance equation. Here two time series (xt and yt) represent the time courses of an interacting miRNA-mRNA pair. T samples is used to delay xt. Ux and Uy are the means of their respective time series, and their are N samples of each time series.

$$r_{xy}(T) = \frac{1}{N-1} \sum (x_t - T - U_x)(y_t - U_y)$$

In the cross-correlation function, O_{xx} and O_{yy} are the variance of the respective time series.

$$r_{xy}(T) = \frac{O_{xy}(T)}{\sqrt{O_{xx}(0) O_{yy}(0)}}$$

This equation can be further reduced to only show the variance between the two time series. This information is from chapter 7 of the Signal Processing Course [211].

$$r_{xy}(0) = \frac{O_{xy}}{O_x O_y}$$

quickDMap

quickDMap will generate a heatmap which is compatible with *quickDendro* output. The heatmap is colour coded to represent magnitude of change over time. *grDevices::colorRampPalette* is used to display genes values.

quickDendro

quickDendro will create a dendrogram from the filtered miRNAs and mRNAs. This function uses *ggdendro::ggdendrogram* to form the dendrogram.

quickFuzz

quickFuzz visualizes clusters created by *createClusters* or *createClusters2*. Using *Mfuzz::mfuzz.plot2*, fuzzy cluster plots are generated [152]. These clusters display different temporal behaviours based on the changes of the number of genes in common between the SDEGs, per time point and the WikiPathways. Intensity of colour represents the levels of fit a pathway has to the temporal behaviour, from highest to lowest: red, orange, yellow, purple. Colours of the plots can be altered. If input from *createClusters2* is used each line is a gene instead of a pathway. Also, if s mode of analysis is used mRNAs and miRNAs

can be visualised one at a time.

quickHClust

quickHClust will create line plots for each gene found in the selected cluster. These plots will have a grey line representing data points and a red smooth spline overlaying the data points. This function uses *stats::cutree, dplyr::inner_join, dplyr::filter, ggplot2::geom_line, ggplot2::geom_smooth* and *ggplot2::facet_wrap* [204, 210].

quickMap

quickMap will create a heatmap of the miRNA-mRNA pairs filtered with the *matrixFilter* function. The pairs will be ordered and numbered based on descending correlation scores.

quickNet

quickNet plots filtered miRNA-mRNA interactions which have been formatted into an *igraph* friendly format by *makeNet*. Pink nodes are miRNAs and blue nodes are mRNAs, and edges are colour coded based on the correlation between the miRNA-mRNA interactions. This function uses *igraph::V*, *igraph::E*, *grDevices::colorRampPalette* and functions from *R* base package *graphics*, including: *graphics::par*, *graphics::plot*, *graphics::legend* [199, 212].

quickPathwayTC

quickPathwayTC displays all genes filtered from the *matrixFilter* function. Each gene is plotted along the time course, and gene expression values are scaled. Users can define a threshold and genes which are higher or lower than the threshold at any time point are highlighted with *gghighlight* [213].

quickReg

quickReg displays regression analysis between a selected miRNA and mRNA that are predicted to interact. OR and 95% CI are pasted as subheadings. The following *R stats* functions are used here: *stats::confint.default* and *stats::coef*.

OR measures an exposure (i.e. miRNA behaviour) and an outcome (i.e. mRNA behaviour). If $OR = 1$ then there is no association, if $OR > 1$ then there is a positive association and if $OR < 1$ there is a weak association. An $OR < 1$ may also indicate a negative relationship, which is what we expect with true miRNA-mRNA interactions. This technique has been performed before for time matched miRNA-mRNA time series analysis in *Jayaswal et al (2009)*, and this idea is used here [158].

There are four components needed to calculate an OR and CI.

- a = no change in miRNA and no change in mRNA
- b = change in miRNA and no change in mRNA
- c = no change in miRNA and change in mRNA
- d = change in miRNA and change in mRNA

$$OR = \frac{a.d}{b.c}$$

$$CI = exp(log(OR) + / - \frac{Z\alpha}{2} . \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$Z\alpha$ = critical value parameter which is calculated by the regression model.

quickTC

quickTC creates a plot displaying a single miRNA-mRNA pair over the time course. This is a line plot, which can be interpolated and scaled. The correlation score will be pasted as a subheading. *FreqProf::approxm* is used here.

quickTCPred

quickTCPred predicts the regression between data of a selected gene (miRNA or mRNA). The genes predicted values are based on binding partners which were selected during creation of a linear regression model by the *linearRegr* function. A mRNA being targetted

by multiple miRNAs or a single miRNA targetting multiple mRNAs can be explored here. R^2 values and P values are pasted as subheadings.

reduceWiki

reduceWiki is best used once a pathway of interest has been selected, either by time course enrichment analysis or temporal cluster analysis. *reduceWiki* will retrieve all the genes associated with the pathway of interest.

returnCluster

returnCluster will retrieve information about any cluster of interest from *createClusters*. Users can specify the threshold of fitness score to be used as a barometer for retrieval of pathways. Fewer pathways will be retrieved if the fitness score threshold is very high. The default is 0.5, and users can define their own threshold between 0 and 1.

savePlots

savePlots stores results from *enrichWiki* in the working directory. A plot will be made for each time point (c mode of analysis) or for each gene type, and time point (s mode of analysis). Plots can be saved as png, jpeg, svg or tiff images.

significantVals

significantVals looks through the list of nested data frames (c analysis) or list of lists of nested dataframes (s analysis) made by the *genesList* function to remove genes which are not deemed ”significantly differentially expressed”. It is advised to include a DE result which represents confidence, and that should be used here, along with a user defined significance threshold. Each nested data frame (time point) will be analysed independently of other time points. This function should be skipped if using data from a non-pairwise based DE method.

startObject

startObject will make a MultiAssayExperiment (MAE) to store the input miRNA and mRNA data [196]. MAEs are the standard objects within *TimiRGeN*. Most dataframes and ma-

trices are stored as assays, lists are stored as metadata and S4 objects are stored as experiments [196]. This allows for cleaner global enviroments.

turnPercent

turnPercent creates a percentage matrix out of the matrix created by *wikiMatrix*. The final row (total gene number) is used for normalisation. This makes a more valid input for *createClusters* or *createClusters2*.

wikiList

wikiList downloads current version of species specific wikipathway information. Unlike *dloadGmt*, WikiPathways data is downloaded as a large list, and each listed pathway is attached to a list of characters, which represent gene IDs associated to the pathway. This function uses *rWikiPathways::listPathways* to download the pathways and *rWikiPathways::getXrefList* to get the gene IDs [124].

wikiMatrix

wikiMatrix uses lists from *wikiList* and *eNames* to generate a matrix which identifies, how many genes are in common between the SDEGs found in each time point, and the pathways. Columns are pathways, and rows are time points, and a final row is added, which represents the total number of genes in each pathway.

wikiMrna

wikiMrna finds genes in common between the WikiPathway of interest and input mRNAs.

2.4 Summary

Overall I have created a novel *R/ Bioconductor* package to integrate, analyse and generate small detailed networks of miRNA-mRNA interactions from big longitudinal multiomic datasets. This package was successfully accepted onto the *Bioconductor* repository. Being a *Bioconductor* tool provides several benefits. Firstly, it means I will receive regular updates on how my tool performs on multiple operating systems (windows, mac and linux) and I will learn if there are potential bugs to be fixed by email. Also, there is a prestige

attached to *Bioconductor* packages, as such it will be more readily trusted by users. Furthermore, there is a strong community of developers who can help if issues arise, not to mention the *Bioconductor* core team consists of many talented individuals who can provide aid and advice if needed. So far this package has had hundreds of downloaded since its acceptance. This tool was the basis of a first author original paper which was published in *Bioinformatics* journal [117].

CHAPTER 3

CHONDROGENESIS DATA ANALYSIS TO FIND MIRNA-MRNA INTERACTIONS

3.1 Background

My PhD was funded by the Dunhill Medical Trust as a collaboration between Newcastle (David Young) and East Anglia (Ian Clark). One of the focuses of the grant was to identify miRNA regulators of developing cartilage. My contributions to this aim would be: a) use a previously generated longitudinal miRNA-mRNA expression dataset to identify miRNA-mRNA interactions for further investigation, b) generate GRNs and kinetic models based on validation data and c) use the kinetic model to make predictions and to organise complex miRNA-mRNA interactions *in silico*.

3.1.1 Cartilage

Muscles move fluidly because the ends of bones are attached to a particular type of extracellular matrix known as cartilage (specifically Hyaline cartilage). Cartilage is a complex mesh of collagenous and protein-saccharide constructs created and maintained by chondrocytes; unique cells which are sparsely scattered throughout the cartilaginous extracellular matrix (ECM) [214, 215, 216]. Along-side chondrocytes, cartilage consists of wa-

ter, inorganic ions, proteoglycans, glycoproteins, glycosaminoglycans (GAGs). In Hyaline cartilage, these components help to create an elastic, compressible and shock absorbing tissue that has important roles in limb formation, skeletogenesis and maintenance of health and function in numerous body parts [215, 216, 217]. Hyaline cartilage and can be found in ribs, the trachea and at the articular ends of long bones; the latter is more specifically called articular cartilage [218, 219]. Articular cartilage is a connective tissue which is smooth and lubricated and it allows for frictionless conformal changes and compression for stress bearing for the surrounding bones. Articular cartilage is able to compress and distribute stress across the tissue, which protects the underlying bone from damage [220].

Bio-mechanics of articular cartilage is complex and relies on many factors to contribute to the function of cartilage, such as a high hydrostatic pressure which is achieved by having a 70-80% water content. During load bearing events, liquid in the cartilage is pushed out slowly, giving way to the more rigid dry sections of the cartilage which is resilient to compression, and after the event, water comes back into the cartilage, restoring its elastic and fluid qualities. Many of these properties can be attributed to important anabolic cartilaginous proteins such as COL2A1 and ACAN.

COL2A1 creates large tough collagen fibers which keeps the chondrocytes in place within the cartilage and provides resilience [221]. Each collagen fiber consists of three collagen protein wrapped in a tight tri-mer conformation that is maintained by cross-links [222, 223, 224, 225]. Non-collagen groups can link together different collagen fibers to create and even more denser and resilient ECM, such a COMP, decoring, lumican and fibromodulin [226].

ACAN is a protein-saccharide construct, specifically a proteoglycan. ACAN molecules can absorb and dissipate impact from external stresses on synovial joints. An ACAN molecule is made up of 3 main domains (G1, G2, G3) and in between G2 and G3 many GAGs (keratin sulfate and chondroitin sulphate chains) are covalently bound to ACAN to create negatively charged structures which attract positively charged H ions of water molecules. The G1 domain of ACAN proteoglycans are attached to hyaluronate which creates rows

of ACAN proteoglycans which can work together to create a resilient structure [227, 227, 228].

Cartilage related diseases

Cartilage is aneural and avascular, and these characteristics means the chondrocytes must largely maintain themselves, and the cartilaginous ECM by controlling catabolic and anabolic procedures within the cartilage for remodelling and homeostasis. During the ageing process the cartilage degrades, and this is most commonly seen after 40 years of age in humans. This condition is known as osteoarthritis (OA) and it can vary in severity. Patients with mild symptoms may be able to get by with habitual changes but those with more severe symptoms may require pain relief drugs or joint replacement surgery [229]. OA is a primary cause of pain and disability for many individuals. Several factors can directly or indirectly impact the progress and severity of OA including: gender, obesity, diabetes, heart disease, injury, occupation and genetics. This makes OA a complex condition, not to mention that different joints in the same individual may be effected by OA at different severities. However, statistics do support ageing to be the primary contributing factor, for example in the United States, over a third of all over 65s have OA, which has been predicted to contribute annually to 3.4 to 13.2 billion dollars in direct and 10.3 billion dollars in indirect costs from OA [230, 231, 232, 233]. Western countries have carried out most epidemiological studies of OA so the global picture is unclear, however it is estimated that around 0.6% of all disability-adjusted life-years (DALYs) and roughly 10% of all musculoskeletal conditions are a result of some form of OA [234].

During OA the catabolic processes outweigh the anabolic ones, leading to a progressive loss of articular cartilage. Molecular changes during the early stages of OA includes the progressive loss of ACAN and COL2A1. Also, OA chondrocytes increase expression of catabolic proteins such as matrix metalloproteinase proteins (MMPs), e.g. MMP1, MMP8, and MMP13 which enzymatically degrade proteins like COL2A1. Also the a disintegrin and metalloproteinases attached to type 1 thrombospondin motifs (ADAMTS) family of proteins also increase in expression e.g. ADAMTS5 which degrades ACAN. The increase in ECM catabolic proteins is what causes the erosion of the cartilagenous ECM.

3.1.2 Chondrogenesis

Chondrocytes develop in a process called chondrogenesis. Fibroblast like mesenchymal stem cell (MSCs) condense and undergo a multi-step process where chondroprogenitor cell differentiate into rounded chondrocytes [235, 236]. This is an important process for limb bud development, skeletogenesis and for cartilage development [235].

The master regulator of chondrogenesis is SOX9. Its transcription is initiated by TGFB signalling and SOX9 promotes expression of COL2A1 and ACAN [113**?**]. Mutations in SOX9 are known to cause Campomelic Dysplasia, Acampomelic Campomelic Dysplasia and Pierre Robin Sequence, the foremost is a severe skeletal malformities and the latter two are milder skeletal deformities [237, 238, 239]. SOX9 is expressed early in chondrogenesis and accumulates to higher levels during the process. Partner proteins help SOX9 to function and enhance chondrogenesis such as FOXC, FOXP and FOXF [240, 241]. Other proteins such as SOX5 and SOX6 increase SOX9s ability to bind to target DNA and function as a transcription factor. Without the SOX-trio, pro-chondrogenic signal would be insufficient to trigger differentiation [242, 243]. Other transcription effectors like RUNX1 and GLI also enhance transcription of pro-chondrogenic genes [244, 245]. Another function of SOX9 is to keep antagonistic transcription factors such as RUNX2 and WNT signalling factors at lower levels to maintain chondrogenic gene expression. This will slow chondrocytes from differentiating further [246, 247].

SOX9 protein undergoes post translational modifications to perform its functions as master regulator of chondrogenesi. This includes phosphorylation at amino acids Serine (S)64 and S181. Phosphorylation occurs by Cyclic AMP dependent protein kinase A (cAMP-PKA) and ROCK1 [248, 249]. SUMOylation also affects SOX9 and can occur after phosphorylation [250]. Several other proteins are known to affect SOX9 stability, localisation and effectiveness such as SIRT1 increasing SOX9 activity by deactylation of NfKB, an antagonist of SOX9 activity [251].

SOX9 gene expression is tightly regulated by multiple pathways (Figure 3.1). For example the SMAD pathways which is controlled by TGFB induces *SOX9* gene expression [252].

Hypoxia inducing factor HIF1a promotes *SOX9* gene expression by transactivation and promoter binding [253]. Other pathways which induce *SOX9* activity include HEDGEHOG, BMP and FGF. On the other hand, pathways which repress *SOX9* gene expression includes WNT and NOTCH signalling [254, 255, 256]. Interestingly certain pathways can have multiple affects on *SOX9* expression, for example TGFB signalling also promotes RHoA and its effector ROCK1, which is thought to inhibit *SOX9* gene expression [257].

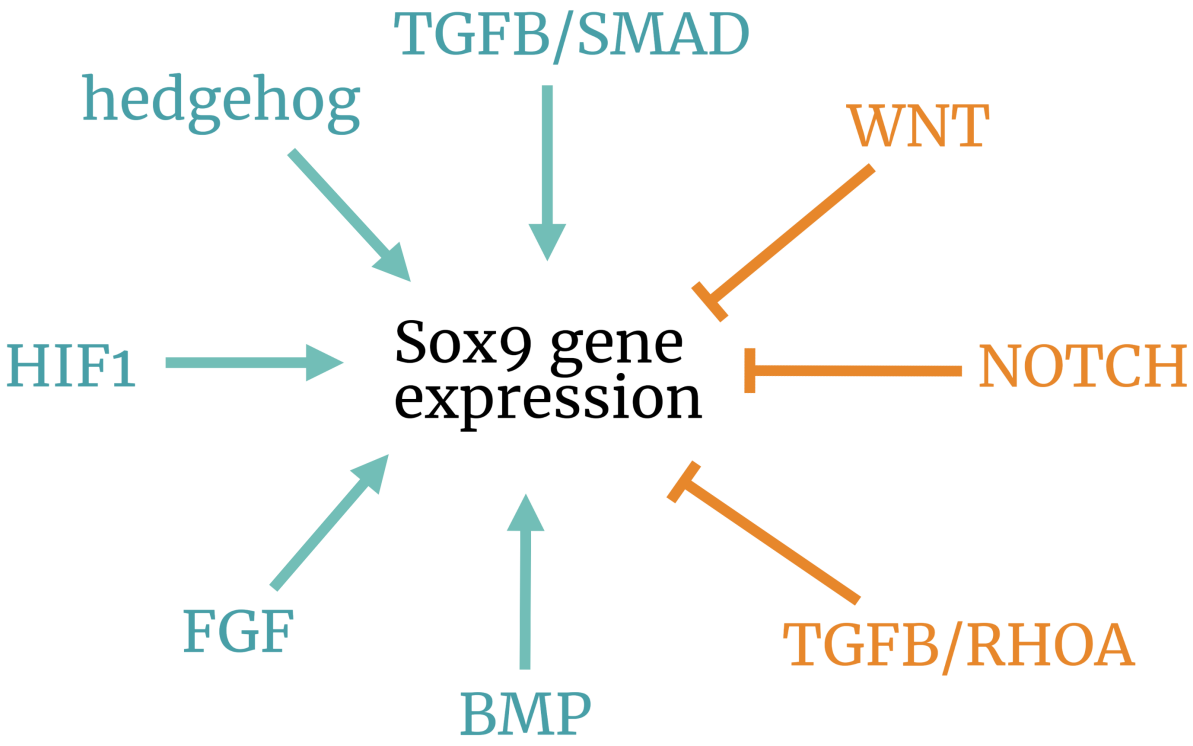


Figure 3.1: **Pathways that regulate *SOX9* expression.** Mind map shows some of the pathways that induce (light blue) or reduce (orange) *SOX9* expression.

Hypertrophic chondrocytes

Articular chondrocytes stay in a quiescent state. Here, they do not undergo proliferation and maintain a consistent expression profile to provide the cartilaginous ECM with homeostasis. However, the expression profile of chondrocytes change, diverting them to a less stable and more cartilage degenerative phenotype. This process is known as hypertrophy [258]. Hypertrophic changes are hallmarks in developing OA. Understanding the gene

expression changes from normal chondrogenesis and hypertophic chondrocytes will lead to better a better understanding of of OA [254, 259]. It should also be mentioned, hypertrophy is a normal and important process in bone growth.

Hypertrophic chondrocytes can be classified by the changes in molecular activity. Cells undergoing this process have reduced gene expression of pro-chondrogenesis markers such as *COL2A1, ACAN, SOX9, SOX5, SOX6* mRNAs, and in contrast pro-hypertrophic markers increase, such as *MMP10, MMP1, ADAMTS5, COL10A1, TNFa* mRNAs [221, 260, 261, 262, 263]. Notably, many bone promoting transcription factors are promoted in hypertrophic chondrocytes, such as RUNX2, VEGFA, HDAC4 and BMP2 [261, 264, 265, 266]. Other regulatory factors such a non-coding RNAs also change in expression during chondrogenesis, hypertrophy and OA [267].

microRNAs regulate chondrogenesis

There are many miRNAs that promote or deter chondrogenesis. In mammalian models, knock out (KO) DICER mutations lead to severe malfunctions in limb formation during embryogenesis; indicating miRNAs do regulate chondrogenesis. Also KO DICER lead to abnormalities in a growth plate and subsequent endochondral ossification during skeletogenesis [268, 269]. Many miRNAs are involved in the regulation of metabolic pathways which promote chondrogenesis, such as *miR-140-5p*. Its function is necessary for chondrogenesis, and maintaining cartilage. Miyuki et al (2010) found that *miR-140-5p* KO mice had skeletal defects and showed signs of early on-set OA [91].

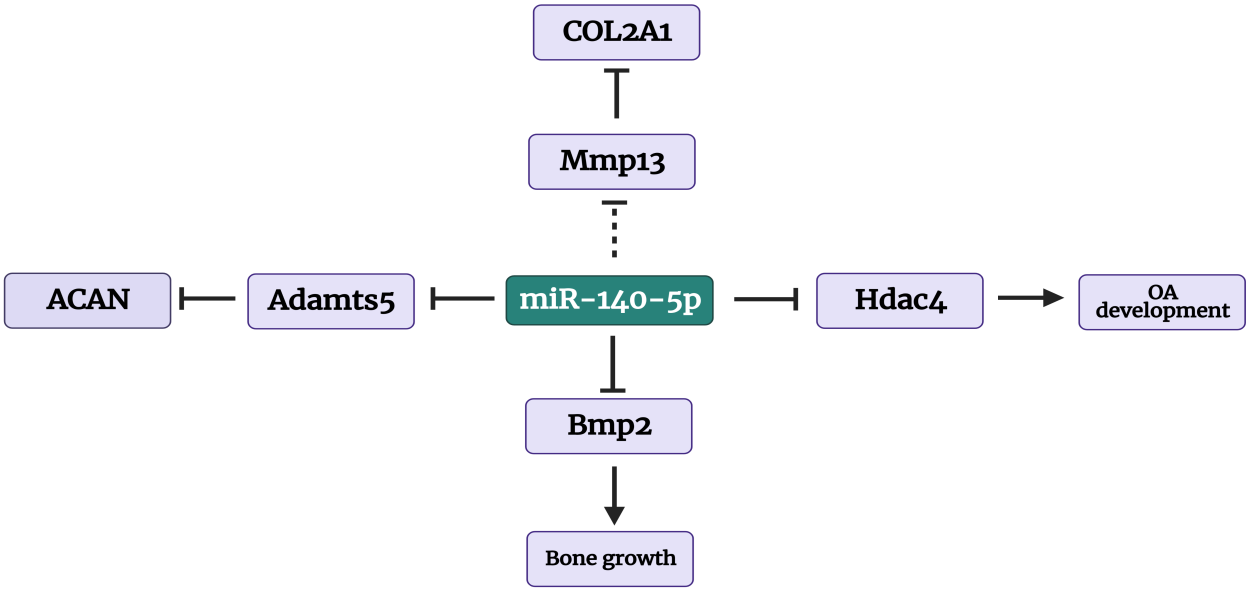


Figure 3.2: **Target genes and affects on chondrogenesis.** Mind map shows target mRNAs which are regulated by miR-140-5p. Dotted lines indicate predicted activity and solid lines represent reported activity. <—means activation and |—means inhibition. |– – means predicted inhibition, without evidence in the literature.

To highlight the complexity involved in understanding the role a miRNA, I will briefly describe some of the functions of *miR-140-5p* in the context of chondrogenesis. *miR-140-5p* has been shown to target *ADAMTS5* for degradation and is predicted to target *MMP13*. Both are catabolic ECM genes that are highly active during OA conditions and directly contribute to the decay of articular cartilage [91, 106, 270]. *miR-140-5p* also targets *HDAC4*, a deacetyl transfase gene which has been shown to have some involvement in OA development [271, 272]. Furthermore, *miR-140-5p* targets genes which promote hypertrophy, such as *BMP2* [273]. Furthermore, understanding the regulation of miRNAs by transcription factors is just as complicated, and SOX9 is a good example for this. Though several important *miR-140-5p* targets are highlighted in this section, in reality dozens of mRNAs regulated by *miR-140-5p* and many of these likely regulate chondrogenesis/ hypertrophy.

SOX9 regulates and is regulated by multiple miRNAs during chondrogenesis. This includes SOX9 being directly responsible for the transcriptional activation of *miR-140-5p* [274]. SOX9 has also been found to transcriptionally repress the activity of some miR-

NAs such as *miR-29a*; an antagonist of NFkB and collagen activity (see subsection 2.2.4) [180, 275, 276]. On the other hand many miRNAs target *SOX9* mRNA for repression. *miR-101* targets *SOX9* mRNA, and *miR-101* downregulation decreases activity of IL-1B induced degradation of COL2A1 and ACAN [277]. *miR-145-5p* also targets *SOX9* mRNA. During chondrogenesis, *miR-145-5p* levels decrease, and during hypertrophy *miR-145-5p* levels increase. Making *miR-145-5p* a potential drug target to prolong cartilage health [278].*SOX9* is also modulated by miRNAs before chondrogenesis. *SOX9* found in MSCs is targetted for degradation by *miR-459-3p* [279].

Other pro-chondrogenesis genes are also effected by miRNAs, such as *miR-194* targetting *SOX5* mRNA. *miR-194* is downregulated during chondrogenesis but is upregulated during OA-like conditions. The reduction of SOX5 activity in these OA-like conditions can contribute to a reduction in the abundance of newly formed anabolic ECM genes. Other miRNAs have been reported to target anti-chondrogenesic genes, such as: *miR-320-c* targetting *ADAMTS5*, *miR-125b* targetting *ADAMTS4* and *miR-27c* targetting *MMP13* [280, 281, 282]. These miRNAs are potential drug targets for intervention in OA.

Understanding miRNAs will help in understanding conditions like OA

A major goal of musculoskeletal research is to generate new chondrocytes from MSCs. The new chondrocytes can generate cartilage for the patients, and may be a novel treatment for OA. Another, aim is to identify drug targets for pain relief which could lead to increased joint function in OA patients. These tasks require a great amount of understanding the process of chondrogenesis, including the transcriptional regulation of important genes such as *SOX9* by miRNAs. Investigations into the intricate affects miRNAs have during chondrogenesis can provide guidance for which miRNAs/ genes to target for therapy. However, identification of the specific miRNAs is made difficult because a single miRNA can regulate many genes and a single gene can be regulated by multiple miRNAs. This is exemplified in mir-ome studies which have found over 18,000 and over 34,000 miRNA-mRNA interactions in HEK293 and hepatoma cells respectively [87, 283]. To add to the complexity the function of a single miRNA can be vastly different in different tissues e.g. *miR-140-5p* has alternate targets within brain e.g *ADAM10* [284]. Investigations

into how miRNAs affect chondrogenesis can be made more efficient if computational work accompanies experimental work.

3.1.3 *In silico* analysis of chondrogenesis

Chondrogenesis is a complex system and to gain novel knowledge from it several computational and systems biology approaches are used. This involved analysing a longitudinal miRNA-mRNA dataset with the *TimiRGeN R* package and analysing network output in *PathVisio* to identify interesting miRNA-mRNA interactions to further study.

Chondrogenesis dataset

I received a longitudinal miRNA-mRNA microarray based dataset from my collaborators in the Young group in Newcastle [113]. This data was gathered from human bone marrow MSCs of seven adult donors aged between 18 and 25 years old. Cells were expanded in a monolayer culture using a MSC growth medium. Chondrogenic culture was used to resuspend the MSCs, and the culture was replaced every 2-3 days, until 14 days. This culture contained several reagents for chondrogenic differentiation including 10 ng/ml of TGFB3, 100 mg/ml of sodium pyruvate, 100 nM of dexamethasone, among other reagents, which are fully described in *Barter et al (2015)* [285]. Total RNA and miRNA were extracted. Illumina whole-genome expression array HT-12 V4 profiled the total RNA samples and Exiqon miRCURY LNAmicroRNA Array was used to profile miRNA samples. Array technologies measured mRNA and miRNA expression along different time points. mRNA data was measured as quadruple at D0 and D14 of chondrogenesis and at singular at intermediate time points: D1, D3, D6 and D7 of chondrogenesis. miRNA expression data was measured as duplicates at each of the time points. The miRNA and mRNA datasets were normalised and pairwise DE was performed with *limma* [146]. The zero time point was contrasted against all other time points for pairwise analysis. log2FC and adjusted P value results from each time base DE analyses were taken forward for further investigation with *TimiRGeN*. The experiments performed by my collaborators worked, as seen by certain chondrogenic genes being highly enriched throughout the time course e.g. *COL2A1*, *SOX9*, *ACAN*, *hsa-miR-140-5p* and *hsa-miR-140-3p*.

3.2 Results

3.2.1 Processing and DE analysis on the chondrogenesis dataset

Raw mRNA data and miRNA data were normalised using the *lumi* and *EximiR* packages respectively [286, 287]. Normalised samples were visualised by PCA plots to see their spread (Figure 3.3).

Using *limma* functions: *makeContrast, constasts.fit, eBayes topTable (BH method)* genes which had an adjusted P value of < 0.05 were classed as SDEGs. Using the zero time point data as the denominator, SDEGs were filtered for each time point. In a chronological order (D1, D3, D7, D10, D14) 3293, 4430, 6049, 4915, 7800 SDEGs were found in the mRNA data. A total of 11562 unique SDEGs were among all of the analyses. Log2FC and adjusted P values were extracted from each time point for the 11562 SDEGs, even if at a particular time point a gene was not a SDEG. I.e. gene X could be classed as a SDEG at only D1 and be filtered for, and gene Y could be classed as a SDEG at every time point, and be filtered for in the same way as gene X. The same procedure was used for the miRNA data and found 48, 127, 210, 210, 217 SDEGs for each time point. A total of 314 unique miRNA SDEGs were found. Log2FCs and adjusted P values for these 314 miRNAs were extracted from each of the DE analyses. The mRNAs (11562) and miRNAs (314) were into *TimiRGeN* as dataframes. Results from mRNA and miRNA DE are shown by volcano plots (Figures 3.4-3.5). miRNAs were generally not as highly expressed or highly differentially expressed as mRNAs, so in contrast the the mRNA volcano plot, fewer miRNAs were highlighted.

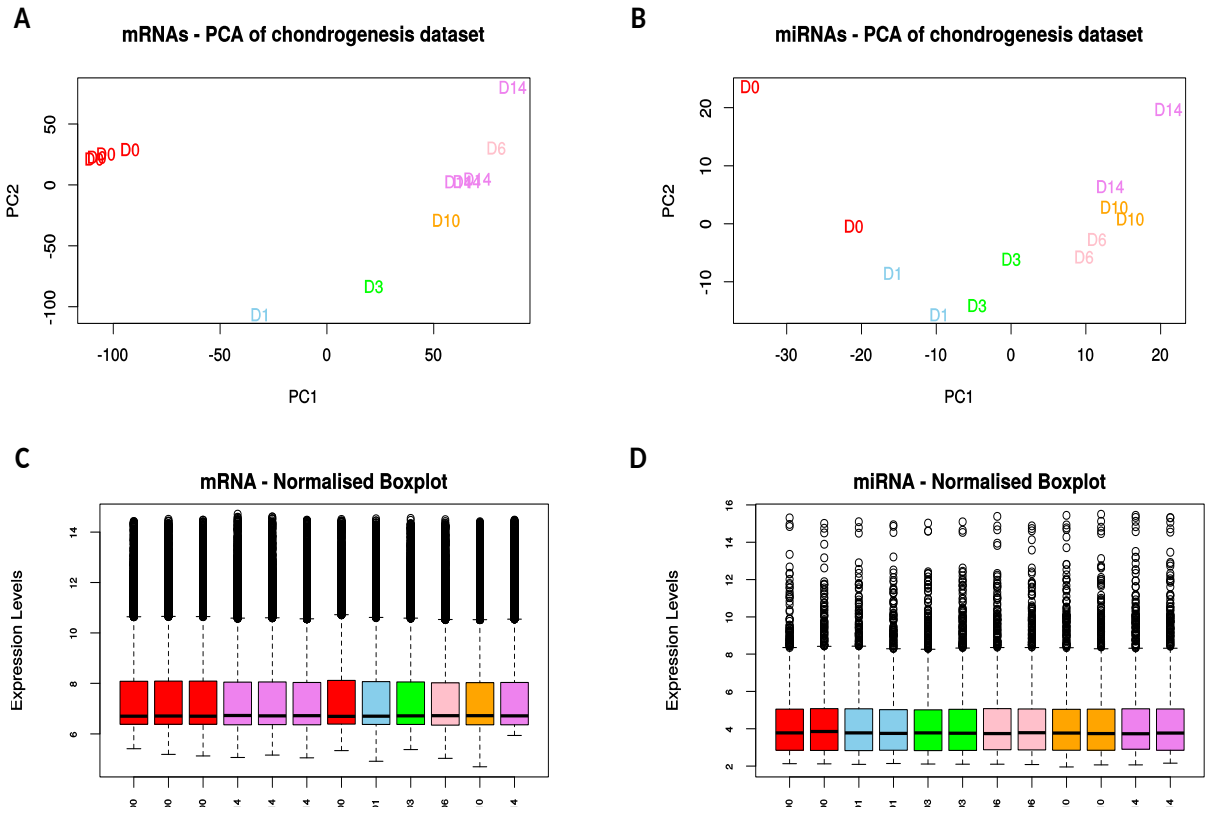


Figure 3.3: **PCA and boxplots showing normalised miRNA and mRNA samples**. miRNA and mRNA samples from the chondrogenesis dataset were analysed by PCAs to show the distance between samples, **A)** for mRNAs and **B)** for miRNAs. Boxplots also show the effect of normalisation on the data, **C)** for mRNAs and **D)** for miRNAs. Samples: D0, D1, D3, D6, D10, D14 are respectively colour coded as: red, blue green, pink, orange, purple.

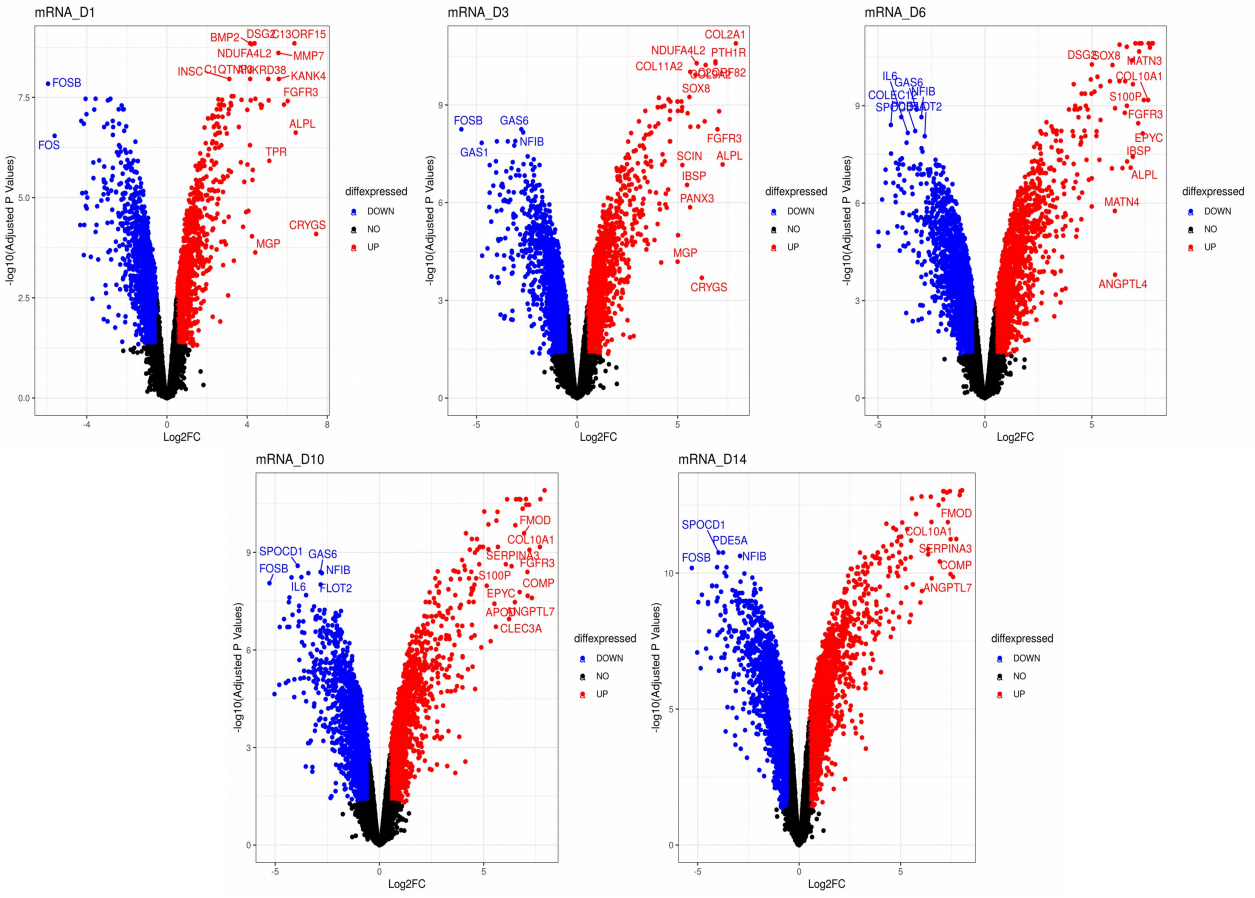


Figure 3.4: **Volcano plots showing DE mRNAs at each time point.** Up (red), down (blue) and non-significantly DE (black) genes across five pairwise analyses, each using the zero time point as the denominator. The cut-offs for genes to be highlighted in these volcano plots are: less than -0.5 log2FC (blue), more than +0.5 log2FC (red), and an adjusted P value of less than 0.05. Results from the DE analyses: D1/D0, D3/D0, D6/D0, D10/D0 and D14/D0 are shown.

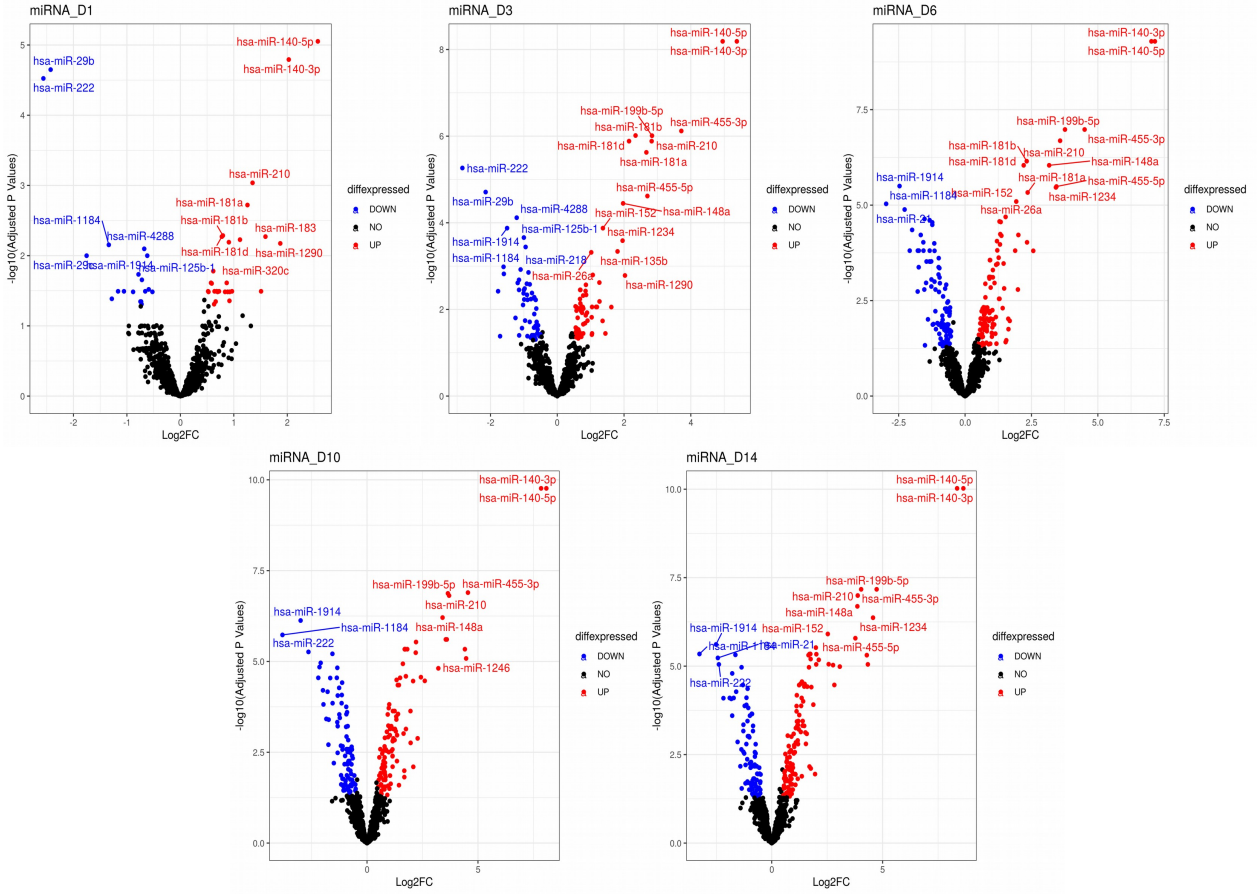


Figure 3.5: **Volcano plots showing DE miRNAs at each time point.** Up (red), down (blue) and non-significantly DE (black) genes across five pairwise analyses, each using the zero time point as the denominator. The cut-offs for genes to be highlighted in these volcano plots are: less than -0.5 log2FC (blue), more than +0.5 log2FC (red), and an adjusted P value of less than 0.5. Results from the DE analyses: D1/D0, D3/D0, D6/D0, D10/D0 and D14/D0 are shown.

3.2.2 TimiRGeN analysis of the chondrogenesis dataset

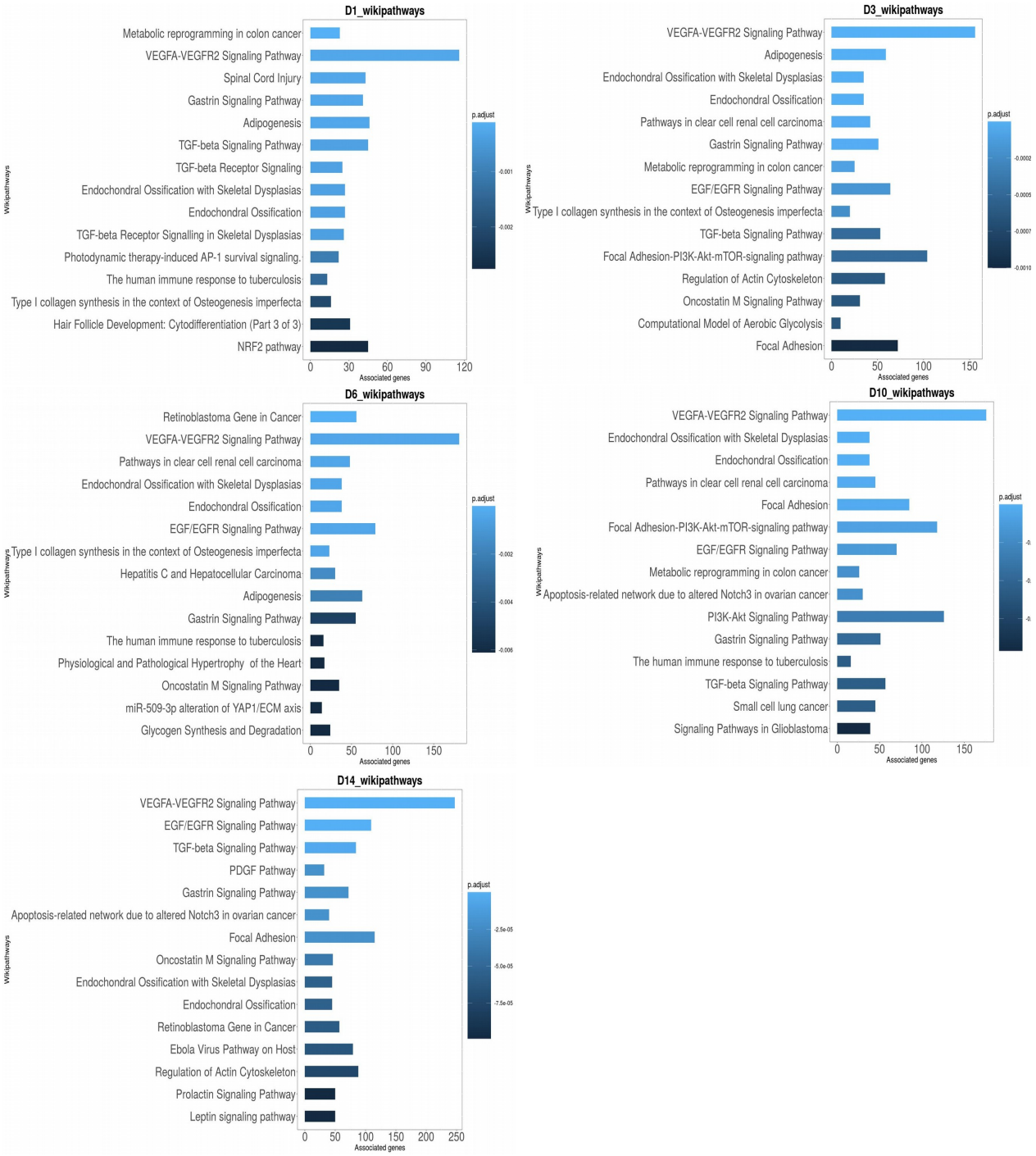


Figure 3.6: **Chondrogenesis dataset analysed by pathway enrichment using *TimiRGeN***. Barplots shows enrichment for each time point. In order; D1, D3, D6, D10 and D14 of chondrogenesis. The lighted the shading, the lower the confidence (adjusted P values).

The chondrogenesis data was analysed using the combined mode of the *TimiRGeN R* package. SDEGs were organized in a way so that each gene is listed with the time points where the genes had an adjusted P value of < 0.05 (see subsection 2.3.2). Time point based pathway enrichment found several pathways which were consistently upregulated during the time course, including VEGFA-VEGFR2 Signalling Pathway, TGF-beta Signaling Pathway, Endochondral Ossification and Apidogenesis (Figure 3.6).

I decided to further investigate the TGF-beta Signaling Pathway because TGFB3 was one of the inputs for chondrogenic differentiation. Analysis with the *R* package found a total of 88 potential miRNA-mRNA interactions were found to regulate the TGF-beta Signaling Pathway over the 14 day time course. These interactions were found by filtering for interactions that had been mined in at least two of the three databases (see subsection 2.1.2). Correlation was not used for filtering because it was a long time course and the averaged correlations could have masked early/ late occurring miRNA-mRNA interactions. Scaled time course plotting revealed the three most positively changing miRNAs to be, in order: *hsa-miR-140-5p, hsa-miR-199b-5p* and *miR-455-5p* (Figure 3.7). *Barter et al (2015)* identified all three to be of interest and focused on *hsa-miR-140* and *hsa-miR-455* [113]. Due to the volume of data that needed to be plotted, *Cytoscape* was used to display all the miRNA-mRNA interactions filtered from the *TimiRGeN R* package and further network analysis was performed in *PathVisio* (Figures 3.8-3.9).

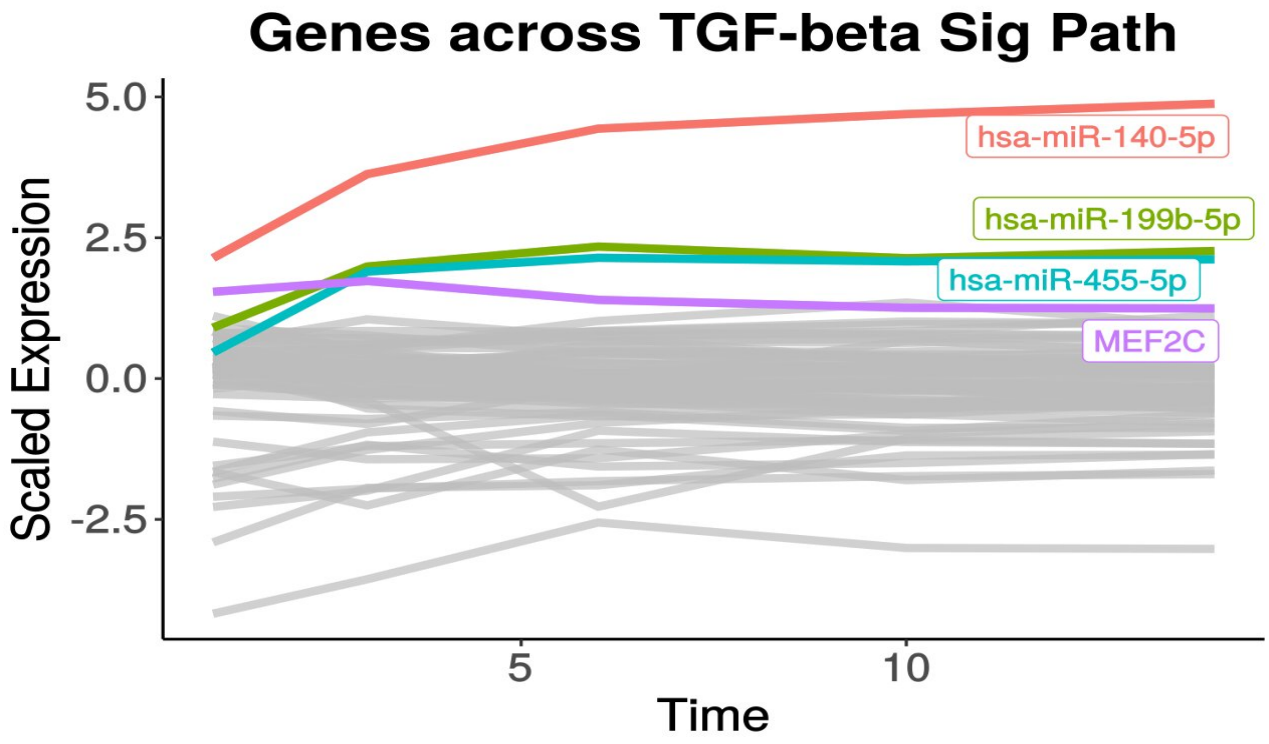


Figure 3.7: **Most positively changing filtered genes.** Filtered genes from the TGF-beta Signalling Pathway were scaled and those which passed the threshold of 1.5 at any point of the 14 day time course are highlighted. This found (in order of highest to lowest) *hsa-miR-140-5p*, *hsa-miR-199b-5p*, *hsa-miR-455-5p* and MEF2C to be the most positively changing genes in terms of magnitude.

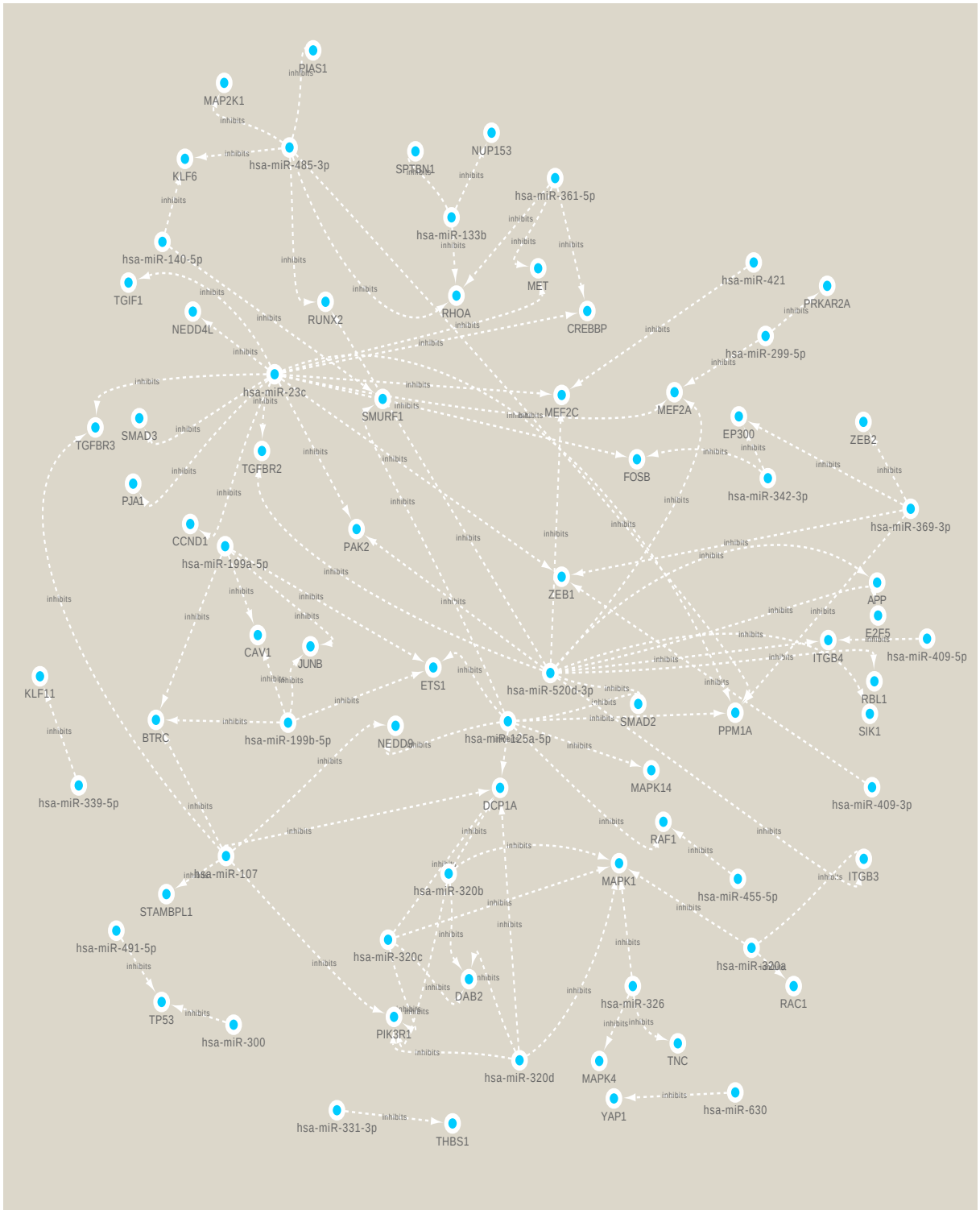


Figure 3.8: **miRNA-mRNA interactions exported to *Cytoscape*.**

Further investigation in *PathVisio*

The miRNA and dynamic data information was exported out of *R* and overlaid onto the TGF-beta Signaling Pathway within *PathVisio*. This created a dynamic miRNA-integrated signalling pathway, which showed how the miRNAs could be influencing the TGF-beta Signaling Pathway during chondrogenesis. Several miRNA-mRNA interaction were of interest, however based on the results from Figure 3.7, *hsa-miR-199b-5p* was the focus because it was a highly positively changing miRNA and novel in chondrogenesis research. *miR-199b-5p* had four predicted mRNA interactions from the TBF-beta Signalling Pathway: *BTRC*, *CAV1*, *ETS1* and *JUNB*, though none had particularly interesting correlations with *hsa-miR-199b-5p*. I used *SkeletalVis*, a portal with many skeletal research related transcriptional datasets that have been analysed by DE, to investigate each of these potential *hsa-miR-199b-5p* targets [288]. Out of the four genes, I found *CAV1* to be the most consistently negatively differentially expressed gene in chondrogenesis studies. Thus, *miR-199b-5p-CAV1* was selected to be the central interacting pair for further computation work. *hsa-miR-199a-5p* was also identified for further study because it is a homologue of *hsa-miR-199b-5p*, so it was likely to have the same targets during chondrogenesis, and may even have a more active role because *hsa-miR-199a-5p* was more highly expressed than *hsa-miR-199b-5p*, though *hsa-miR-199b-5p* had a greater magnitude of change over the time course. Directly downstream of CAV1 was RHoA. Output from *TimiRGeN* also predicted *hsa-miR-361-5p* and *hsa-miR-485-3p* to target *RHoA*.

Analysis with *SkeletalVis*

SkeletalVis was used to identify if *BTRC*, *CAV1*, *ETS1* and *JUNB* are significantly downregulated during chondrogenesis, specifically during time course MSC differentiation studies or time matched chondrocytes-MSC studies [288]. Table 3.1 shows the results of the analysis. *SkeletalVis* contained two suitable studies which can be found from GSE109503 and GSE18394 [289, 290]. The first study was a 28 day human MSC to chondrocyte differentiation, though the DE performed was not zero denominator pairwise, and rather was stepwise. The other study was a 28 day cow chondrocytes-MSC contrast study, with time matched MSCs and chondrocytes for comparisons.

Study	Context	*BTRC*	*CAV1*	*ETS1*	*JUNB*
GSE109503	D1MSC/D0MSC	-0.11	-0.99	1.24	0.42
GSE109503	D3MSC/D1MSC	0.06	-0.221	0.01	-0.3
GSE109503	D7MSC/D3MSC	0.15	-1	-0.84	0.15
GSE109503	D14MSC/D7MSC	0.02	-0.323	0.13	0.09
GSE109503	D21MSC/D7MSC	0.14	0.04	0.31	0.53
GSE109503	D21MSC/D14MSC	0.11	0.366	0.17	0.43
GSE18394	D28chon/D0MSC	0.45	-0.207	-1.61	0.11
GSE18394	D28chon/D28MSC	0.34	0.365	-0.97	0.68
GSE18394	D28MSC/D0MSC	0.10	0.57	-0.64	-0.57

Table 3.1: **Log2FC values of *miR-199b-5p* targets from other MSC studies.** Stepwise or time-matched DE analyses of two other time course chondrogenesis studies from *SkeletalVis*. The log2FC values for each gene (*BTRC*, *CAV1*, *ETS1*, *JUNB*) is given across the analyses.

Contrasts shown in Table 3.1 indicated *CAV1* to be the most consistently downregulated of the four predicted *miR-199b-5p* targets.

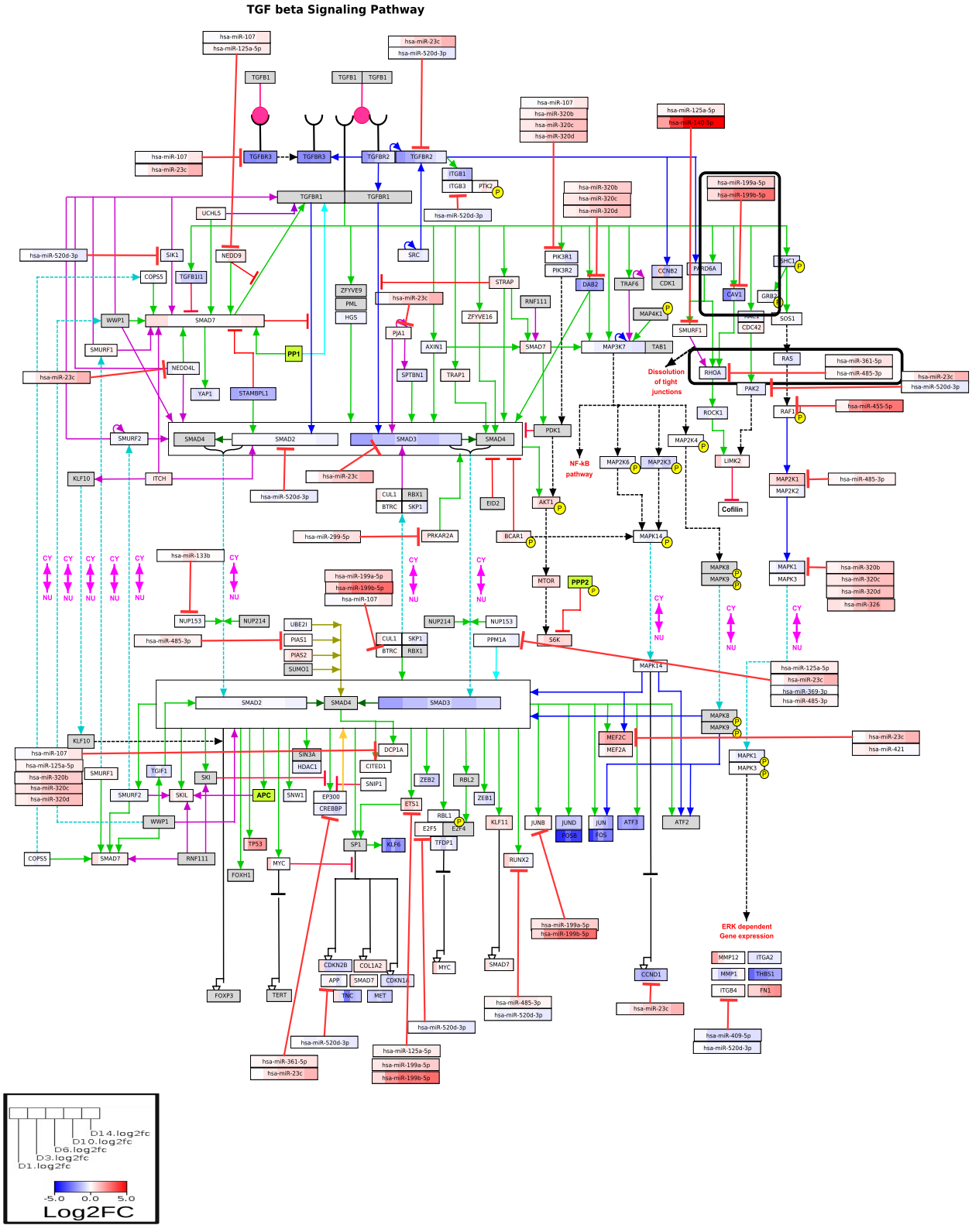


Figure 3.9: **miRNA integrated dynamic TGF-beta Signalling Pathway.** The TGF-beta

Signalling Pathway was analysed in *PathVisio*. Dynamic data (log2FC) values of the 14D time course are added in a chronological order, and a colour code from -5 to +5 shows how each gene in the pathway changed over the time course. miRNAs were added, and their interactions with predicted mRNA targets were drawn by red |—. Black squares were drawn around sections of the image to show sections used for bottom-up GRN construction.

The miRNA data was checked to see if the miRNAs had a high expression level. Table 3.2 displays the miRNAs. It can be seen that *hsa-miR-361-5p* changed marginally during the course of chondrogenesis and *hsa-miR-485-5p* was consistently lowly expressed. *hsa-miR-199a-5p* was highly expressed and *hsa-miR-199b-5p* increased greatly. So out of the four miRNAs of interest, *hsa-miR-199a-5p* and *hsa-miR-199b-5p* were the most likely to be contributing to chondrogenesis.

miRNA	D0	D1	D3	D6	D10	D14
hsa-miR-199a-5p	10.54	11.17	11.38	11.58	11.70	11.88
hsa-miR-199b-5p	5.42	6.33	8.26	9.18	9.06	9.45
hsa-miR-361-5p	6.96	7.18	7.21	7.62	7.65	7.73
hsa-miR-485-3p	4.93	5.14	5.00	5.68	5.56	5.25

Table 3.2: **Target miRNA expression levels.** Normalised expression levels from microarray analysis show how each of the candidate miRNAs are expressed over the time course. These miRNAs may target CAV1 mRNA or RHoA mRNA. These are relative numbers because they come form array technologies. These miRNAs were found from *TimiRGeN-PathVisio* exploration.

Looking at the mRNA targets (Table 3.3), *RHoA* did not show as much variance during chondrogenesis as *CAV1*. *CAV1* decreased in the early stages of chondrogenesis, and then increased to reach a steady state at D6. This implies *CAV1* may have anti-chondrogenic contributions during the early phase of chondrogenesis and may be required for homeostasis during the later phase. Therefore, predicted miRNA regulators of *CAV1*,

hsa-miR-199a-5p and *hsa-miR-199b-5p* may be pro-chondrogenic in function as they increase over the whole time course, though this is only speculation without validatory work. A possible early regulation by the miRNA targets may also explain why *hsa-miR-199b-5p-CAV1* and *hsa-miR-199a-5p-CAV1* had uninteresting correlations over the time course.

mRNA	D0	D1	D3	D6	D10	D14
CAV1	13.21	9.83	9.89	11.68	11.19	11.21
RHoA	12.89	12.76	12.56	12.29	12.11	12.22

Table 3.3: **Target mRNA expression levels.** Normalised expression levels from microarray analysis show how each of the candidate mRNAs which could be targeted by miRNAs (*hsa-miR-199a-5p, hsa-miR-199b-5p, hsa-mir-361-5p, hsa-miR-485-3p*) for regulation during chondrogenesis that were found from *TimiRGeN - PathVisio* exploration.

miR-199a-5p was found to directly target *CAV1* in the context of lung fibrosis, lung inflammation and apidogenesis [291, 292, 293]. *miR-199b-5p* was also reported to directly target *CAV1*, however it seems *miR-199a-5p/ miR-199b-5p-CAV1* are novel interactions in the context of chondrogenesis [294]. *miR-361-5p* and *miR-485-3p* have not been reported to directly target *RHoA*. Given this information, a *miR-199a/b-5p-CAV1* centred chondrogenesis kinetic model may be of interest. However, the mechanism which underpins TGFB regulation of CAV1 activity during chondrogenesis must be established. For this the RHoA/ROCK1 signalling pathway seemed a logical starting point because it is downstream of CAV1 activity in the Wikipathway explored and the RhoA/ROCK1 system is documented well.

3.2.3 Sequence analysis of the miRNA-mRNA interactions

miRNA-mRNA interactions: *miR-199b-5p-CAV1, miR-199a-5p-CAV1, miR-361-5p-RHoA, miR-485-3p-RhoA* were further analysed by sequences specificity between the miRNA seed sites and the 3'UTRs of the mRNA targets.

miR-199a-5p's sequence is $CCCAGUGUUCAGACUACCUGUUC$ and its seed site is $UGUGACC$. It is predicted to bind to one sequence on the 3'UTR of *CAV1*, which is at nt positions 1573-1579, which is $ACACUGG$.

miR-199b-5p shares a similar sequence to *miR-199a-5p*, with 2 mis-matched, but the same seed site. The sequence is $CCCAGUGUUUAGACUAUCUGUUC$ and it's seed site is
$UGUGACC$. It also bind to positions 1573-1579 of the 3' UTR of *CAV1*.

miR-361-5p's sequence is $UUAUCAGAAUCUCCAGGGGUAC$ and its seed site is $AGACUAU$. *miR-361-5p* has one potential binding region on the 3'UTR of *RHoA* which is at nt positions 930-937, which is $UCUGAUA$.

miR-485-3p's sequence is $GUCAUACACGGCUCUCCUCUCU$ and its seed site is $ACAUACU$. *miR-485-3p* has one potential binding region on the 3'UTR of *RHoA* which is at nt positions 250-256, which is $UGUAUGA$.

3.3 Methods

3.3.1 Processing and analysis

The microarray data was normalised with the *lumi R* package. The *VST* (Variance Stabilizing Transform) method was used from normalisation. VST used bead level expression to calculate normalising weights [286]. Gene names were identified using the *lumiHumanAll.db* package [295]. Normalised data underwent standard processing with the *limma R* package to obtain pairwise DE for each time point. The denominator was always the 0 time point. This was carried out for the mRNA and miRNA data. Since the intermediate time points of the mRNA dataset were at $n = 1$, one-tailed T-tests were performed by the *limma* package, rather than a two-tailed T-test. This may have resulted in some biased, towards to 0 time point/ denominator, as this was at $n = 4$. The miRNA data was imported into *R* and Cy3-Cy5 channels were analysed using the *limma* and *EximiR* was used to perform spike-in normalisation [146, 287]. This was used because spike-in probes were

detected in the raw data. DE for the miRNA data was also performed using *limma*, and the 0 time point was used as the common denominator for all analyses.

Results from DE were used as the input for standard analysis with the *TimiRGeN R* package. The combined mode of analysis found the TGF-beta Signalling Pathway to be of interest. Time course analysis identified *hsa-miR-199b-5p* to be the second most highly positively changing miRNA during chondrogenesis. *CAV1* was highlighted as a target of interest because it was significantly down regulated in several chondrogenesis studies found in *SkeletalVis* [288].

3.3.2 Pathway analysis with PathVisio

A MAPP txt file containing miRNA information and a dynamics csv file containing log2FC values of all the genes from the input data were generated. *PathVisio v3.3.0* and several apps were used to create Figure 3.9. The *WikiPathways* app was used to load the TGF-beta Signalling Pathway onto *PathVisio*. Next missing entezgene IDs from the MAPP file and dynamics file were filled in using information from NCBI. Using the *MAPP* app, the miRNAs were loaded into *PathVisio* along with their entrezgene ID annotations. Then the dynamics file was added into *PathVisio* as a dataset. Now the log2FCs for each gene found in the input data and the Wikipathway of interest could be visualised. *PathVisio* allows for multiple data points to be visualised for each gene, which was a useful method to see how gene expression changed over time. Several miRNAs with missing entrezgene IDs had to be manually inserted into the MAPP and dynamics file. The adjusted miRNA entrezgene IDs were used to stop multiple miRNAs from sharing the same IDs. This would have caused bugs in the visualisation as some miRNAs would have been allocated multiple sets of log2FC values.

3.4 Summary

I applied the *TimiRGeN R* package to a chondrogenesis based longitudinal miRNA-mRNA expression dataset. Results from the analysis identified the TGF-beta Signalling Path-

way to be enriched in several time points during chondrogenesis, and that *hsa-miR-199b-5p* was a potential pro-chondrogenic regulatory miRNA. *CAV1* was found as a potential mRNA target of *hsa-miR-199b-5p*, and this was supported by *in silico* investigations with *SkeletalVis* and literature. *hsa-miR-199a-5p* was also found to be of interest because it is a homologue of *hsa-miR-199b-5p*.

Further pathway analysis in *PathVisio* identified RHoA/ROCK1 signalling is downstream of TGFB induced CAV1. *TimiRGeN* also predicted *hsa-miR-361-5p* and *hsa-miR-483-3p* to target *RHoA* mRNA. Following on from these results, it may be interesting to generate a miRNA based chondrogenic kinetic model which may involve the following interactions: *miR-199a-5p-CAV1, miR-199b-5p-CAV1, miR-361-5p-RHoA, miR-483-3p-RHoA*. Though, *miR-361-5p-RHoA* and *miR-483-3p-RHoA* were questionable because their dynamics over the time course were respectively, unvarying and low-level. Overall, through the use of the *TimiRGeN R* package four novel miRNA-mRNA interactions have been found which may regulate chondrogenesis and also serve as the basis for building a kinetic model. *miR-199b-5p-CAV1* was of particular interest for model building and *in vitro* experiments which are shown in Ch4.

CHAPTER 4

MULTI-MIRNA CHONDROGENESIS MODEL

4.1 Background

To identify how *miR-199b-5p* influences chondrogenesis via *CAV1*, I investigated RHoA/ROCK1 activity during chondrogenesis. RHoA/ROCK1 signalling is downstream of CAV1 in the TGFB signalling pathway. Interestingly, CAV1 phosphorylation at Tyrosine (Y)14 has been reported to activate RHoA [296, 297, 298]. Other parts of the TGFB induced RHoA/ROCK1 system have been investigated in this section to establish how the miRNA is affecting chondrogenesis, though much of this information is from non-chondrogenesis research, and so I assume the TGFB-RhoA/ROCK1 system is ubiquitous across cell types.

4.1.1 Biology of RHoA/ ROCK1 signalling

RHoA is a member of the Rho family of small GTPases, and contributes towards chondrogenic maturity and cytoskeletal remodelling. RHoA is activated when it changes into its GTP bound form, from its GDP bound form. Active RHoA phosphorylates protein kinases to affect gene expression [299, 300]. ROCK1 is an effector protein of RHoA, and it phosphorylates downstream kinases to alter gene expression or actin stability. For ex-

ample, RHoA/ROCK1 signalling induces LIM KINASE which then phosphorylates actin depolymerising protein COFILIN, leading to actin stability [301, 302]. Within the context of chondrocytes, RHoA promotes a fibroblast-like cell shape, whereas normal chondrocytes are more rounded. Overexpression of RHoA in chondrocytes inhibits early chondrogenesis and hypertrophy, but ROCK1 has a complex relationship with SOX9 which warrants further investigation [303].

RHoA/ROCK1 regulates SOX9 activity

Inhibition of ROCK1 by inhibitor Y27632 leads to an increase in *SOX9* mRNA [257, 304]. The increase was measured in three different populations of chondrocytes:

- Monolayer of primary cell culture chondrocytes.
- Monolayer of ATDC5 cells.
- 3D micromass culture from limb buds of 11.5 day mouse embryos.

It was reported that ROCK1 negatively regulates *SOX9* mRNA, and the mechanism proposed was that ROCK1 phosphorylates an unknown transcription factor. This theory is strengthened by a luciferase assay showing a portion of the SOX9 promoter region becoming activated during ROCK1 inhibition [257]. Furthermore, inhibition of RHoA by drug Cy3 also lead to an increase in *SOX9* levels in ATDC5 monolayer and HAC cells from patient knee cartilage [257, 305].

However, a later study based in synovium derived MSCs, concluded ROCK1 inhibition by Y27632 leads to a decrease in TGFB1 induced *SOX9* mRNA [306].

More confusing is the effect of RHoA/ROCK1 inhibition on chondrogenic markers other than *SOX9* mRNA. In monolayer primary cells and ATDC5 cells, inhibition of ROCK1 leads to an increase in *L-SOX5*, *SOX6*, *ACAN* and *COL2A1* and overexpression of RHoA lead to a decrease in those four genes [304]. Whereas in the 3D micromass culture, inhibition of ROCK1 lead to a decrease in *L-SOX5, SOX6, ACAN* and *COL2A1* and the latter two genes also had reduced gene expression when ROCK1 was inhibited under TGFB1

stimulated SMSC cells, which were cultured as a monolayer [304, 306]. These results are summarised in Table 4.1.

Study	Culture type	Experiment	*SOX9*	*COL2A1*	*ACAN*	*L-SOX5*	*SOX6*
Woods(2005)	3D micromass	↓ ROCK1	↑	NA	NA	NA	NA
Woods(2005)	ATDC5 monolayer	↑ RHoA	↓	↓	NA	NA	NA
Woods(2006)	ATDC5 monolayer	↑ RHoA	↓	↓	↓	↓	↓
Woods(2006)	ATDC5 monolayer	↓ ROCK1	↑	↑	↑	↑	↑
Woods(2006)	PC monolayer	↓ ROCK1	↑	↑	↑	NA	NA
Woods(2006)	3D micromass	↓ ROCK1	↑	↓	↓	↓	↓
Kumar(2009)	PC monolayer	↓ RHoA	↑	↑	↑	↑	↑
Xu(2012)	SMSC monolayer	↓ RHoA	↓	↓	↓	NA	NA
Xu(2012)	SMSC monolayer	↓ ROCK1	↓	↓	↓	NA	NA

Table 4.1: **Chondrogenesis biomarkers measured in *RHoA/ROCK1* studies.** Results from four studies, including five different chondrocyte cell lines. Decrease/ ↓ or increase/ ↑ of RHoA or ROCK1 lead to an increase or decrease in *SOX9, COL2A1, ACAN*, *L-SOX5* or *SOX6* mRNAs. NA is given if a biomarker was not measured during a study.

A number of possible explanations to these contradicting results have been proposed, including cell culture differences and metabolic links between RHoA/ROCK1 signalling and chondrogenesis biomarkers.

4.1.2 Differences in cell culture methods

One plausible reason could be differences between micromass and monolayer cultures [304].

Differences between culture types: monolayer vs micromass

2D monolayers and 3D micromass culture were generated using different methods and

this could lead to expression differences. This idea was discussed in *Woods and Brier (2006)* and also in *Tew and Hardingham (2006)*, where they identified differences between chondrogenesis in monolayer and micromass cultures [304, 305]. Though, *Xu et al (2012)* showed that chondrogenic markers *ACAN* and *COL2A1* decrease in TGFB1 stimulated SMSCs monolayer culture [306]. Indicating the differences seen in these studies could have resulted from other causes.

Differences between culture types: Culture purity

A further culture issue could also be attributed to cell culture purity levels, meaning some cultures may have high purity as we'd only expect one cell type in their population, whereas other cultures may have a mixed population of cells, and thus have a lower purity. Primary chondrocytes are very pure, ATDC5 cell cultures are also very pure as they are a cell-line and micromass cell cultures are usually a mixed culture. This could have contributed to the contradictory results [304]. The rat SMSCs monolayer culture culture was also a pure cell line [306].

4.1.3 RHoA/ROCK1 regulation of Chondrogenesis

Other than differences between cellular cultures, the gene expression differences could have arisen from the complex relationship between ROCK1 and SOX9.

Difference in regulation: L-SOX5 and SOX6 contribute to SOX9 activity

L-SOX5 and SOX6 are essential for cartilage formation and double KO of these genes are lethal in mice [242]. L-SOX5 and SOX6 can heterodimerise and then bind with SOX9 to form a SOX-trio complex. SOX9s ability to bind to DNA increases in this conformation and has been shown to increase transactivity in regulating *COL2A1*, *ACAN* and *miR-140-5p* which are important markers for chondrogenesis and maintenance of articular cartilage [274, 307, 308]. ROCK1 inhibition in micromass culture lead to decreases in *L-SOX5* and *SOX6* mRNA which could have then contributed to a lowered ability of SOX9 to bind to *COL2A1* and *ACAN* promoters [304]. L-SOX5 and SOX6 activity certainly has an impact

of SOX9 activity, though this was not seen in the ATDC5 monolayer.

Actin polymerisation reduces SOX9 activity

RHoA/ROCK1 activity stabilizes actin polymerisation and this has been linked to reducing gene expression of chondrogenic markers including: *SOX9, L-SOX5, SOX6, COL2A1* and *ACAN* mRNAs in primary caudal sternal (middle of chest) chondrocyte and micromass cells from chicken embryos [309]. Pharmacological interventions to mimic the effects of RHoA/ROCK1 modulation of actin polymerisation were conducted by *Woods and Brier (2006)* who increased actin polymerisation with Jasplankinolide in micromass culture, which lead to an increase in *L-SOX5* and *SOX6* [304]. However, experiments from *Woods et al (2005)* showed that both a reduction of actin polymerisation with Cytochalisin D and an increase in actin polymerisation with Jasplankinolide in micromass culture lead to an increase in *SOX9* [257]. Thus, actin polymerisation affects *SOX9* activity during chondrogenesis but the results makes judging the effects of actin polymerisation on SOX9 activity difficult. Also, it is likely culture type (2D vs 3D) had an impact in these results.

Study	Culture type	Experiment	*SOX9*	*COL2A1*	*ACAN*	*L-SOX5*	*SOX6*
Woods(2005)	3D micromass	stabilisation	↑	↑	↑	NA	NA
Woods(2005)	3D micromass	inhibition	↑	↑	↑	NA	NA
Woods(2006)	3D micromass	stabilisation	↑	↑	↑	↑	↑
Woods(2006)	3D micromass	inhibition	↓	↓	↓	↓	↓
Woods(2006)	PC monolayer	stabilisation	↓	↑	↑	NA	NA
Woods(2006)	PC monolayer	inhibition	↑	↑	↑	NA	NA
Kumar(2009)	PC monolayer	stabilisation	↓	NA	NA	NA	NA

Table 4.2: **Chondrogenic biomarkers measured after alterations in actin stability** Results from three studies, including two different chondrocyte cell lines. decrease/ ↓ or increase/ ↑ of actin stability leads to an increase or decrease in *SOX9, COL2A1, ACAN, L-SOX5* or *SOX6* mRNAs. NA is given if a biomarker was not measured during a study.

Interestingly, *Haudenschild et al (2010)* found that SOX9 phosphorylation by ROCK1

was independent of RHoA/ROCK1 actin remodelling. Co-transfection of LIM KINASE, an actin remodelling protein which is phosphorylated by ROCK1, in SW135 micromass cells showed no significant difference in pSOX9 abundance [302, 310]. Since pSOX9 is required for chondrogenic signal, this means actin polymerisation may only affect *SOX9* expression, and not affect expression of *COL2A1* or *ACAN*.

ROCK1 phosphorylates SOX9

cAMP-PKA phosphorylates SOX9 at S64 and S181, which increased nuclear localisation and is needed for the nuclear import of SOX9 via the importin B-mediated pathway [248, 311, 312]. cGMP-PKII also phosphorylates SOX9 at S181 but this attenuates SOX9 DNA binding properties which may promote hypertrophy during endochondral ossification [313]. ROCK1 is known to directly phosphorylate SOX9 at S181 and inhibition of ROCK1 by Y27632 delays peak pSOX9 levels by 2 days, during chondrogenesis [304, 309, 310]. *Haudenschild et al (2010)* showed ROCK1-mediated pSOX9-181 had increased nuclear localisation, but this diminished with the addition of Y27632 in SW1353 micromass cells. They also identified that ROCK1-SOX9 phosphorylation was part of TGFB signalling and dynamic compression response pathways [310].

ROCK1 transactivates SOX9 by phosphorylating SMAD2/SMAD3

TGFB is a necessary external stimuli for chondrogenesis because it activates *SOX9* transcription via SMAD2/3 signalling; as such TGFB commonly used as a reagent to induce chondrogenic differentiation from MSCs [113, 306, 314]. RHoA/ROCK1 signalling has been found to phosphorylate SMAD3 which then promotes SOX9 transactivation [310, 315]. To add to this, *Xu et al (2012)* found with qPCR experiments, ROCK1 inhibition by Y27632 lead to a decrease in TGFB1 induced pSMAD2 and pSMAD3 in SMSC monolayer [306].

ROCK1 negatively regulates *SOX9* mRNA

It should also be noted again that *SOX9* mRNA levels increase in most experiments after

RHoA/ROCK1 inhibition. The exception was the SMSC monolayer experiments. The assumption behind this increase in *SOX9*, is that here is an unknown *SOX9* transcriptional repressor which ROCK1 activates [257, 304, 305, 306].

TGFB signalling induces miRNAs which may regulate RHoA/ROCK1

miRNAs which regulate the TGFB-RHoA/ROCK1 system have not been well studied and investigating their role can increase our knowledge of TGFB-SOX9 signalling during chondrogenesis. Results from *TimiRGeN* analysis predicted *RHoA* to be targeted by *hsa-miR-361-5p* and *hsa-miR-485-3p*. *hsa-miR-485-3p* expression was increasing over the time course, but was relatively low compared to the other miRNAs. Whilst *hsa-miR-361-5p* expression was stable during the chondrogenesis time course and relatively higher than *hsa-miR-485-3p*'s expression levels (Table 3.2 and Table 3.3). *RHoA* gene expression is gradually decreased along chondrogenesis which was expected because it had been identified as a gene which halts chondrogenic maturity [304]. These miRNAs may regulate *RHoA* during chondrogenesis and thus may have some influence on SOX9 activity.

Also, other proteins, upstream of RHoA/ROCK1 may be regulated by miRNAs. The TGFB signalling pathway identified CAV1 as one such protein, which when phosphorylated mediates RHoA-GDP to RHoA-GTP conversion. *CAV1* mRNA also decreased over chondrogenesis and *TimiRGeN* analysis predicts that *CAV1* mRNA is targeted by *hsa-miR-199a-5p* and *hsa-miR-199b-5p*, and both miRNAs increased in expression levels over the chondrogenesis time course (Table 3.2 and Table 3.3). Interestingly, *CAV1* expression has been researched in the context of chondrogenesis and osteogenesis [316, 317].

Which potential ROCK1-SOX9 mechanisms will be modelled

The complex relationship between ROCK1 and SOX9 would benefit from a kinetic model as a means to capture the mechanisms and make predictions on how RHoA/ROCK1 activity regulates chondrogenesis biomarkers. The size of this model was determined by the level of validation data generated for the model, so not all of the regulatory effects were explored. The selected mechanisms were used to construct a GRN, which was then

modelled. The regulatory mechanisms which were explored by the model were as follows: ROCK1 phosphorylates SOX9, ROCK1 positively regulates *SOX9* by phosphorylating SMAD2 and SMAD3, ROCK1 negatively regulates *SOX9* via an unknown transcription factor, and miRNAs that are predicted to regulate the TGFB-SOX9 system.

Several of the aforementioned regulatory differences could not be included in the model for several reasons. Firstly, the culture differences could not have been tested in lab, and also would be difficult to dissect with a kinetic model. Actin stability was an interesting question but *Haudenschild et al (2010)*, concluded actin regulation by LIM KINASE contributed no significant affect on pSOX9 levels, and this influenced me not to pursue this line of inquiry in a model. L-SOX5 and SOX6 regulation were also an interesting factor, but many of the studies did not report on *L-SOX5* and *SOX6*, in contrast the other chondrogenic biomarkers (*COL2A1, ACAN, SOX9*) were more well documented in this system (Table 4.1 and 4.2), and so there was more knowledge to draw for these genes. Finally, I did not include *hsa-miR-361-5p* and *hsa-miR-485-3p* in the model because validatory results for *hsa-miR-361-5p* (see subsection 4.2.2) indicated it did not greatly affect *RHoA* mRNA and *hsa-miR-485-3p* was assumed to have a marginal affect because it had low expression.

4.2 Results

4.2.1 Gene regulatory networks

A GRN is a blueprint for a kinetic model and a resource to design wet-lab experiments. Here two GRNs are presented. Firstly, a whole GRN which shows all the mechanisms which we wish to explore with a kinetic model. However, what can be modelled is based on the validatory experiments performed (discussed later in this section). A second GRN shows the mechanisms which were modelled. Components of the whole GRN are described below. Both were made using *CellDesigner* [318]. Below I also detail the evidence which supports the topology of the GRNs.

I assumed the mRNA levels to be a proxy to total gene activity (mRNA, protein and phospho-protein), because most of our data was RNA based.

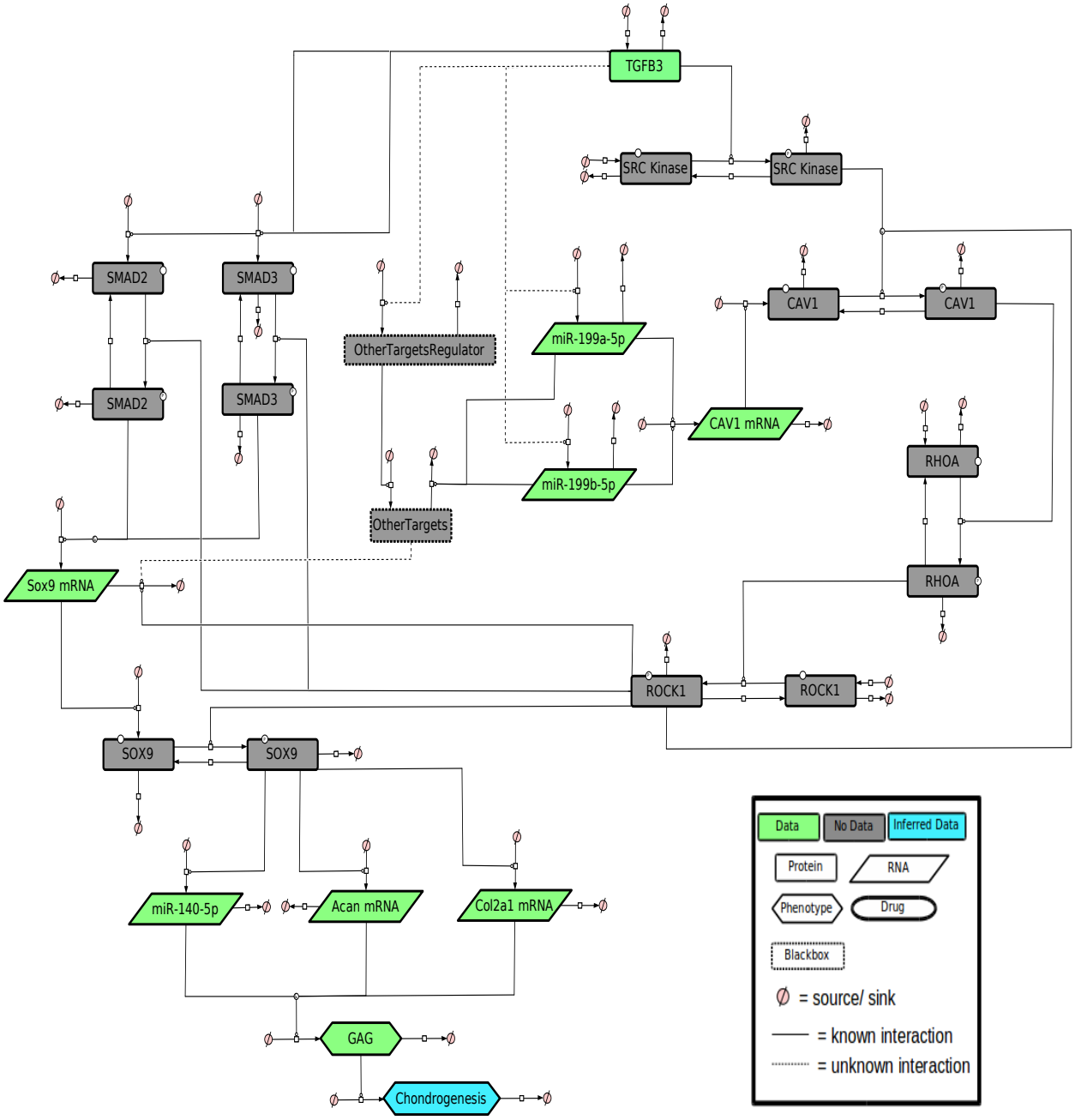


Figure 4.1: **Whole multi-miRNA chondrogenesis model.** This GRN was the foundation of what was modelled, however also contained other species which were not modelled because of data constraints.

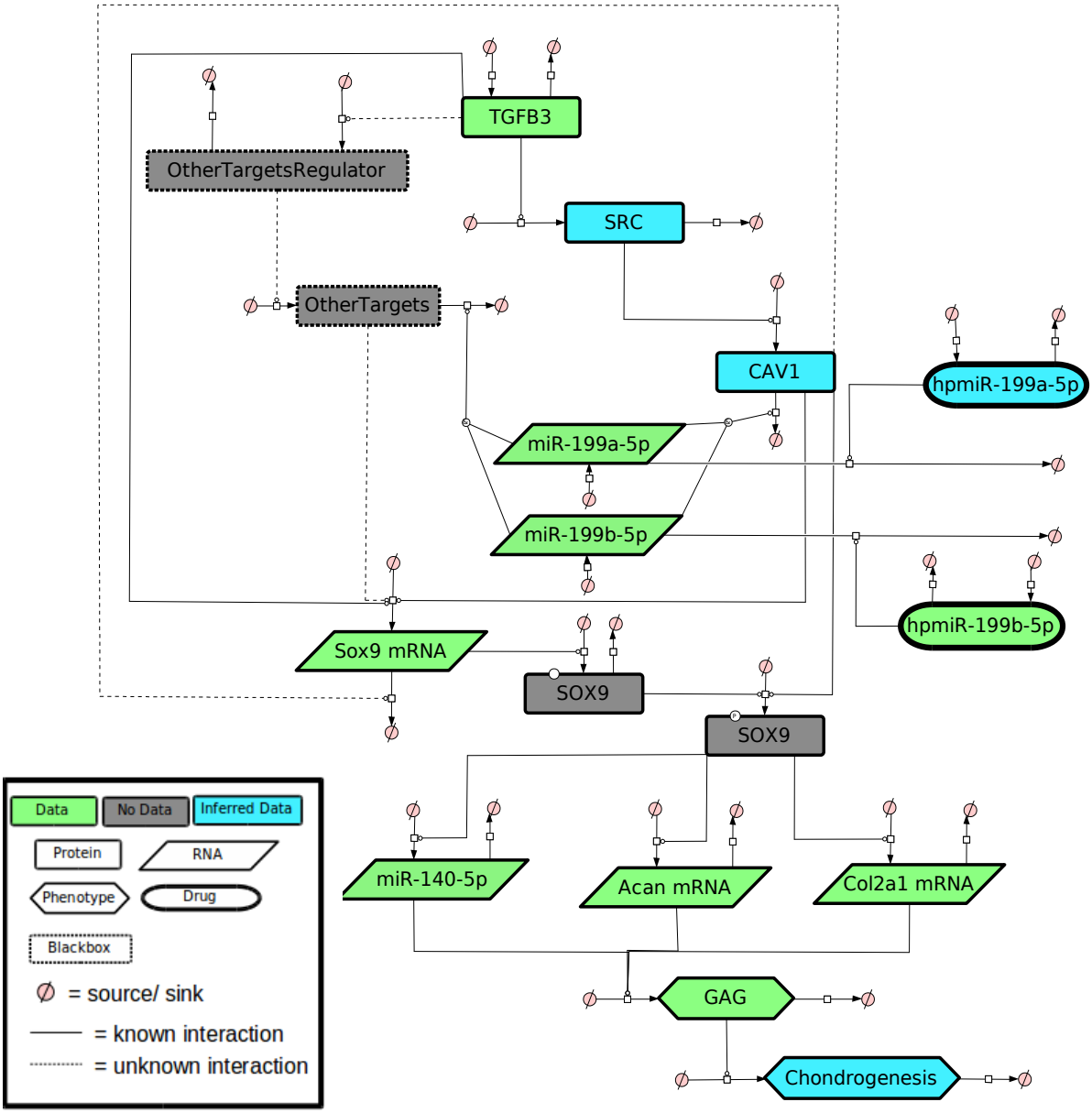


Figure 4.2: **Modelled multi-miRNA chondrogenesis model.** This GRN contains what was modelled in COPASI. Species in the model can be parameterized on all the needed data (green), parametrized based on inferred data (blue), which means we had partial data for that species (e.g. only having mRNA level data (e.g. CAV1) or inferring behaviour based on similar genes (e.g. miR-199a-5p)), or there is no supporting data associated to the species for parameterization (grey).

Several species had to be removed from the whole chondrogenesis model because our validation data did not cover them. This includes RHoA/ROCK1. SOX9 is the exception because ROCK1 and TGFB have alternative regulative affects on *SOX9* mRNA, SOX9 protein and SOX9 phospho-protein. Throughout this model, we assume CAV1 activity is a proxy for RHoA/ROCK1 activity. This assumption helped to speed up the model.

TGFB-SRC kinase

Chondrogenesis begins by stimulation of a number of reagents described in section 3.3, including of TGFB (TGFB3). SRC kinase, a tyrosine kinase from the TGF-beta Signalling Pathway, auto-phosphorylates on Y416 in the presence of TGFB. Interestingly inhibition of SRC kinase by inhibitor SU6656 lead to no RHoA-GTP (active RHoA) formation and inhibition of TGFB also lead to no RHoA-GTP formation [297, 319]. *SRC* kinase decreased in expression level during chondrogenesis, and this is seen in the microarray data produced by our collaborators (Table 4.3) [113]. Also it has been found that SRC kinase inhibits early chondrogenesis, so its downregulation is unsurprising [320].

Gene	D1	D3	D6	D10	D14
SRC	0.19	-0.16	-0.76	-0.64	-0.46

Table 4.3: ***SRC* expression change over chondrogenesis.** Log2FC results from the microarray dataset shows a decrease in the levels of *SRC* during chondrogenesis.

SRC kinase - CAV1

TGFB induced SRC kinase phosphorylates CAV1 at Y14 [297]. *CAV1* is overexpressed in OA conditions so it may have some regulatory functions during chondrogenesis [321].

CAV1 - RHoA

CAV1 activates RHoA-GTP from RHoA-GDP. Mutation Y14A on CAV1 lead to a decrease

in RHoA-GTP. This is because the Y14A mutated CAV1 protein could not be phosphorylated by SRC kinase [297]. Therefore, there is a direct link between TGFB and RHoA activity, via CAV1 and SRC, which has been reported regulate activity RHoA/ROCK1 signalling in cancer and cardiovascular studies [296, 297, 298]. CAV1 is also important during chondrogenesis in chicken limb development [316].

CAV1 - miR-199a/b-5p

CAV1 mRNA is a validated target of *miR-199a-5p* and has been recorded as a target in several biological niches including tissue injury, lung inflammation and apidogenesis [291, 292, 293]. *miR-199b-5p* has been less well studied than its homologue, but it is likely that it targets the same mRNAs due to their similar sequences. *miR-199b-5p* has been recorded to target *CAV1* but never during chondrogenesis [294, 322].

RHoA - ROCK1

ROCK1 is an effector protein of RHoA. RHoA/ROCK1 signalling has been reported to inhibit hypertrophy. RHoA/ROCK1 are also active prior to chondrogenesis starting and decrease in activity over the course of chondrogenesis [304, 323].

ROCK1 - SOX9

As described in subsection 4.1.3, ROCK1 phosphorylates SOX9, induces *SOX9* mRNA via phosphorylation of SMAD2/SMAD3 and negatively regulates *SOX9* mRNA by activating an unknown repressor.

SOX9 - pSOX9

SOX9 is phosohorylated at to make pSOX9. pSOX9 is needed for nuclear localisation and transcription factor activity [248, 312].

pSOX9 - miR-140-5p, Acan mRNA, Col2a1 mRNA

As discussed in subsection 3.1.2, activated SOX9 transcriptionally activates chondrogenic biomarkers: miR-140-5p, *ACAN* mRNA, *COL2A1* mRNA.

miR-140-5p, Acan mRNA, Col2a1 mRNA - GAG

Again as discussed in subsection 3.1.2, GAGs are attached to ACAN molecules to form necessary structures in the Cartilage, so reduction in ACAN would lead to a fall in GAGs. Some GAGs may also be attached to collagen structures, however COL2A1's contribution is indirect, as COL2A1 is needed to maintain healthy cartilage, and thus is needed to maintain GAG levels. Finally, *miR-140-5p* also has an indirect influence on GAG levels as it negatively regulates antagonistic proteins of ACAN and COL2A1.

GAG - Chondrogenesis

I am using GAG levels as a proxy for chondrogenesis in this model, i.e. higher the GAG, the further along the process of chondrogenesis the model is in. This was the only phenotypic level data received, and it was a quantitative measure of chondrogenesis.

4.2.2 Validatory Data from collaborators

Experimental data was generated to test if *hsa-miR-199b-5p* regulates chondrogenesis. Data in this section is from qRT-PCR experiments performed by Dr. Matt Barter in David Young's research group in Newcastle University. Here, hairpins (hp) were used to reduce the levels of *miR-199b-5p* (hpmiR-199), *miR-361-5p* (hpmiR-361) or as a control (hpCon). Hairpins were added at day 0 and the gene expression of chondrogenic biomarkers (Figure 4.3) or predicted miRNA targets (Figure 4.4-4.5) were measured at days 0, 1, 3 and 7.

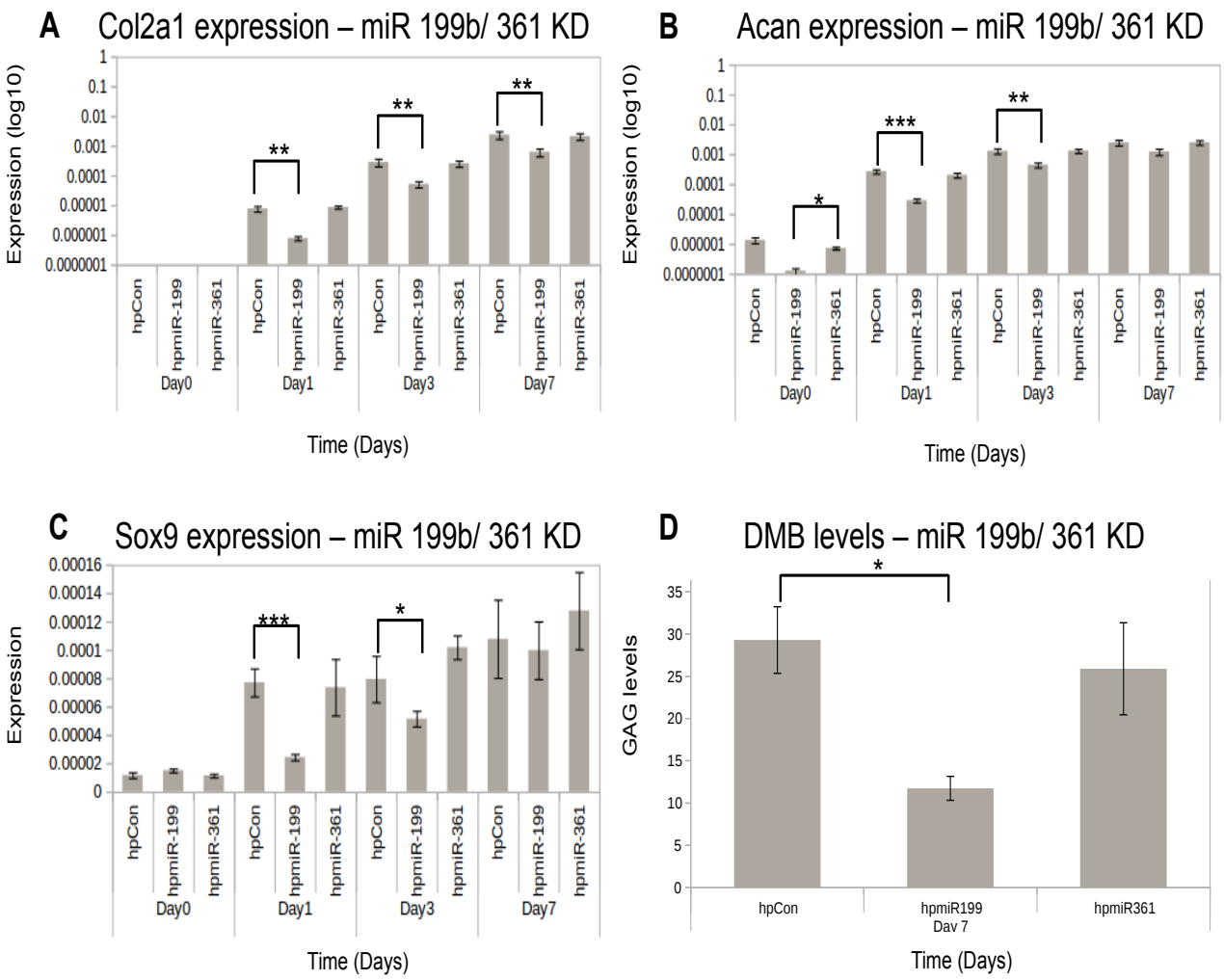


Figure 4.3: **Chondrogenesis biomarker levels after miRNA inhibition. A)** *COL2A1* expression levels significantly decreased after during *hsa-miR-199b-5p* inhibition at each time point after day 0. qRT-PCR results for day 0 were undetectable for *COL2A1* mRNA. **B)** *ACAN* expression levels significantly decreased at days 0, 1, and 3 after *hsa-miR-199b-5p* inhibition. **C)** SOX9 expression levels significantly decreased at days 1 and 3 after *hsa-miR-199b-5p* inhibition. **D)** dimethyl blue staining showed GAG levels significantly decreased after *hsa-miR-199b-5p* inhibition at day 7. Error bars were calculated by standard deviation between 2-3 replicates. Significance was calculated by T-tests between the control and inhibitions. * = $<$0.05, ** = $<$ 0.0001, *** = $<$ 0.00001.

Chondrogenic biomarkers *COL2A1, ACAN* and *SOX9* mRNAs were measured after inhi-

bition of *hsa-miR-199b-5p* or *hsa-miR-361-5p* during Days 0, 1, 3 and 7 of chondrogenesis (Figure 4.3). GAG levels are also measured at day 7. *hsa-miR-199b-5p* was identified to be a pro-chondrogenic regulator. Significant decreases seen in *SOX9* levels may have contributed to decreasing levels of *COL2A1* and *ACAN* levels. The decrease in *ACAN* would likely lead to the significant decrease in GAG levels. *hsa-miR-361-5p* did not show any significant affect on any chondrogenic biomarker, so it is likely not an important regulator of chondrogenesis. For this reason *hsa-miR-361-5p* was not added to the GRNs.

CAV1 is upregulated after *hsa-miR-199b-5p* inhibition and *RHoA* is unaffected by *hsa-miR-361-5p*

CAV1, *HES1* and *JAG1* levels were measured at 0, 1, 3 and 7 days after *hsa-miR-199a-5p* or *hsa-miR-361-5p* inhibition. *CAV1* was tested due to being a predicted target from *TimiRGeN*. *HES1* and *JAG1* were tested due to being known targets of *miR-199b-5p* (Figure 4.4) [122, 324]. Similarly, *RhoA* was a predicted target of *hsa-miR-361-5p*, while *TWIST1* and *VEGFA* were known targets of *miR-361-5p* (Figure 4.5) [325, 326]. Based on results in this subsection, *CAV1* may be a genuine target of *hsa-miR-199b-5p*. Results also indicated that *hsa-miR-361-5p* was not affecting *RHoA* and may not have had any significant effect during chondrogenesis. Interestingly, *hsa-miR-199b-5p* may have some affect on *RHoA*. Other genes such as *HES1* and *VEGFA* may also have been genuine targets of *hsa-miR-199a-5p* and *hsa-miR-361-5p*, though *HES1* expression was also upregulated by *hsa-miR-361-5p* inhibition. *JAG1* did not seem to be a *hsa-miR-199b-5p* target during chondrogenesis, though *JAG1* was reported as a *miR-199b-5p* target [122]. However, *JAG1* may be a target of *hsa-miR-361-5p*. Lastly, *TWIST1* may be a target of *hsa-miR-199b-5p* and *hsa-miR-361-5p* during chondrogenesis. *TWIST1* has been identified as a true target *miR-361-5p*, but it showed a greater response to hsa-*miR-199b-5p inhibition*. It should be noted, qPCR does not prove that *hsa-miR-199b-5p* targets *CAV1*, it only proves that during *hsa-miR-199b-5p* inhibition *CAV1* levels increased.

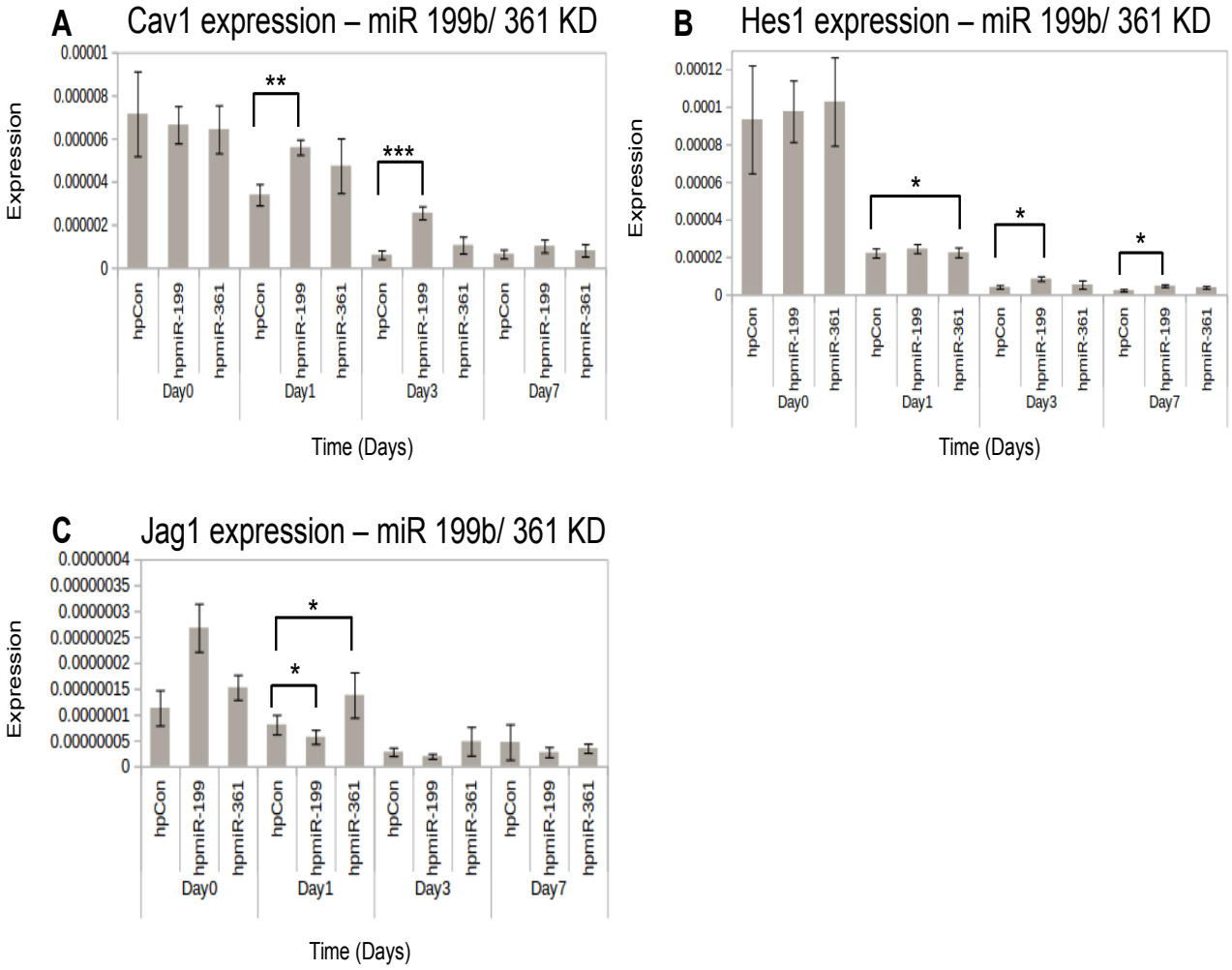


Figure 4.4: **Predicted *miR-199b-5p* targets after miRNA inhibition**. **A)** *CAV1* was significantly upregulated after *hsa-miR-199b-5p* inhibition at days 1 and 3. **B)** *HES1* mRNA was significantly upregulated at day 1 after *hsa-miR-361-5p* inhibition and at days 3 and 7 after *hsa-miR-199a-5p* inhibition. **C)** *JAG1* mRNA was significantly dowregulated after *hsa-miR-199b-5p* inhibition and upregulated after *hsa-miR-361-5p* inhibition at day 1. Error bars were calculated by standard deviation between 2-3 replicates. Significance was calculated by T-tests between the control and inhibitions. * = $<$ 0.05, ** = $<$ 0.0001, *** = $<$ 0.00001.

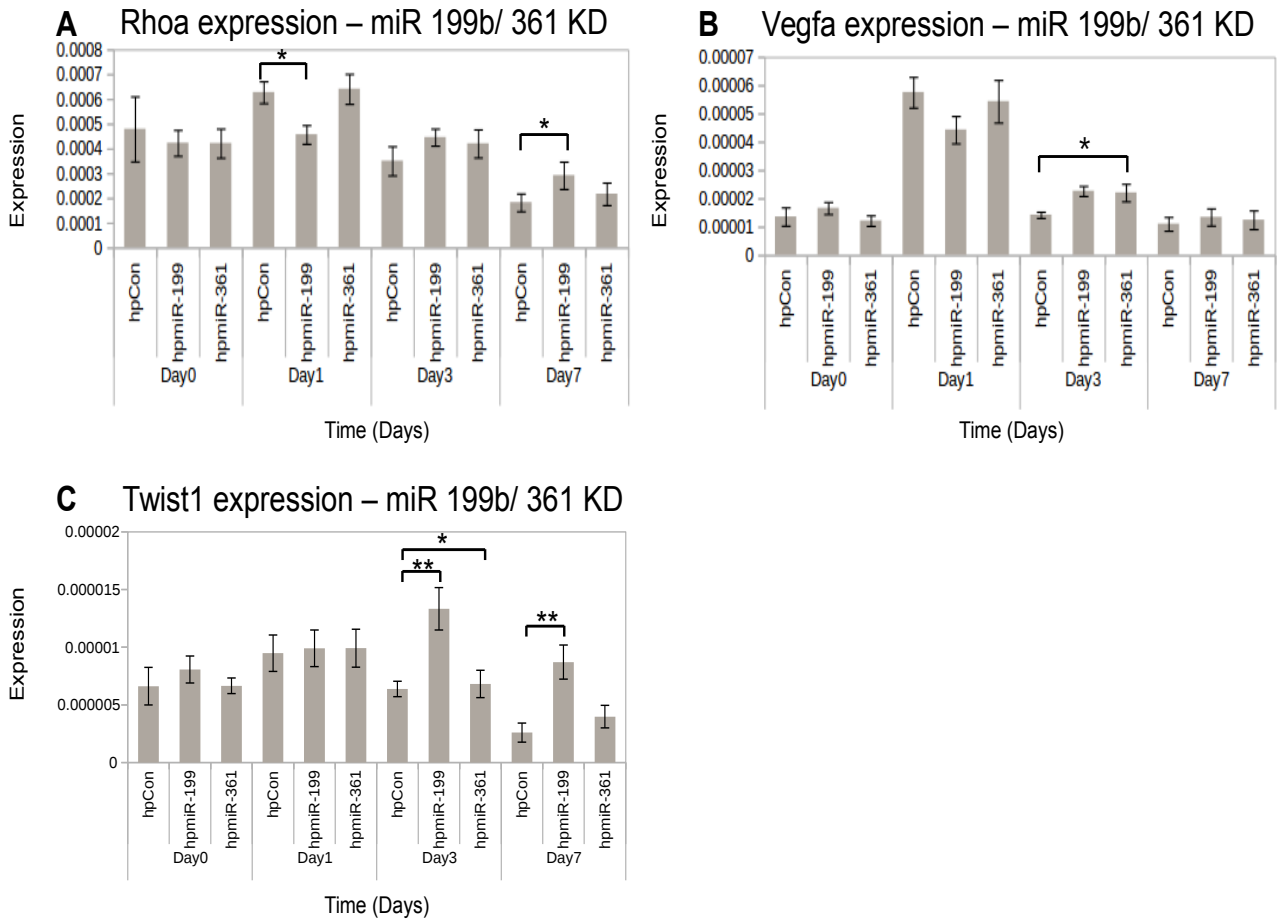


Figure 4.5: **Predicted *miR-361-5p* targets after miRNA inhibition A)** After *hsa-miR-199b-5p* inhibition *RHoA* was significantly downregulated at day 1 and upregulated at day 7. **B)** *VEGFA* was significantly upregulated at day 3 after *hsa-miR-361-5p* inhibition. **C)** *TWIST1* was significantly upregulated at days 3 and 7 after *hsa-miR-199b-5p* inhibition and at day 3 after *hsa-miR-361-5p* inhibition. Standard deviation of 2-3 replicates Error bars were calculated by standard deviation between 2-3 replicates. Significance was calculated by T-tests between the control and inhibitions. * = $<$ 0.05, ** = $<$ 0.0001, *** = $<$ 0.00001.

Though some of these results were significant, it may also be that *hsa-miR-199b-5p* and *hsa-miR-361-5p* are having indirect regulatory affects on these genes. More specific tests, such as luciferase assays, should be carried out to truly confirm that these are direct miRNA-mRNA interactions [5]. That being said, these results are certainly enough to begin kinetic modelling.

4.2.3 Kinetic Modelling

The GRN shown in Figure 4.2 was modelled using *COPASI* [119]. Results from this modelling software were exported into *R* for plotting. After searching on Biomodels, the repository for reproducible kinetic models, it seems that the model generated in the PhD is the first multi-miRNA chondrogenesis model which has validated data associated with it [327].

Model Calibration

Microarray data from *Barter et al (2015)* was used as the calibration data. This is used as a basis to simulate objects within the kinetic model during normal chondrogenesis over the 14 day time course [113]. Overall the model gets the correct trends for all species. Some dynamics are missed in CAV1 and miR-199a-5p, however the general trends are match (Figure 4.6).

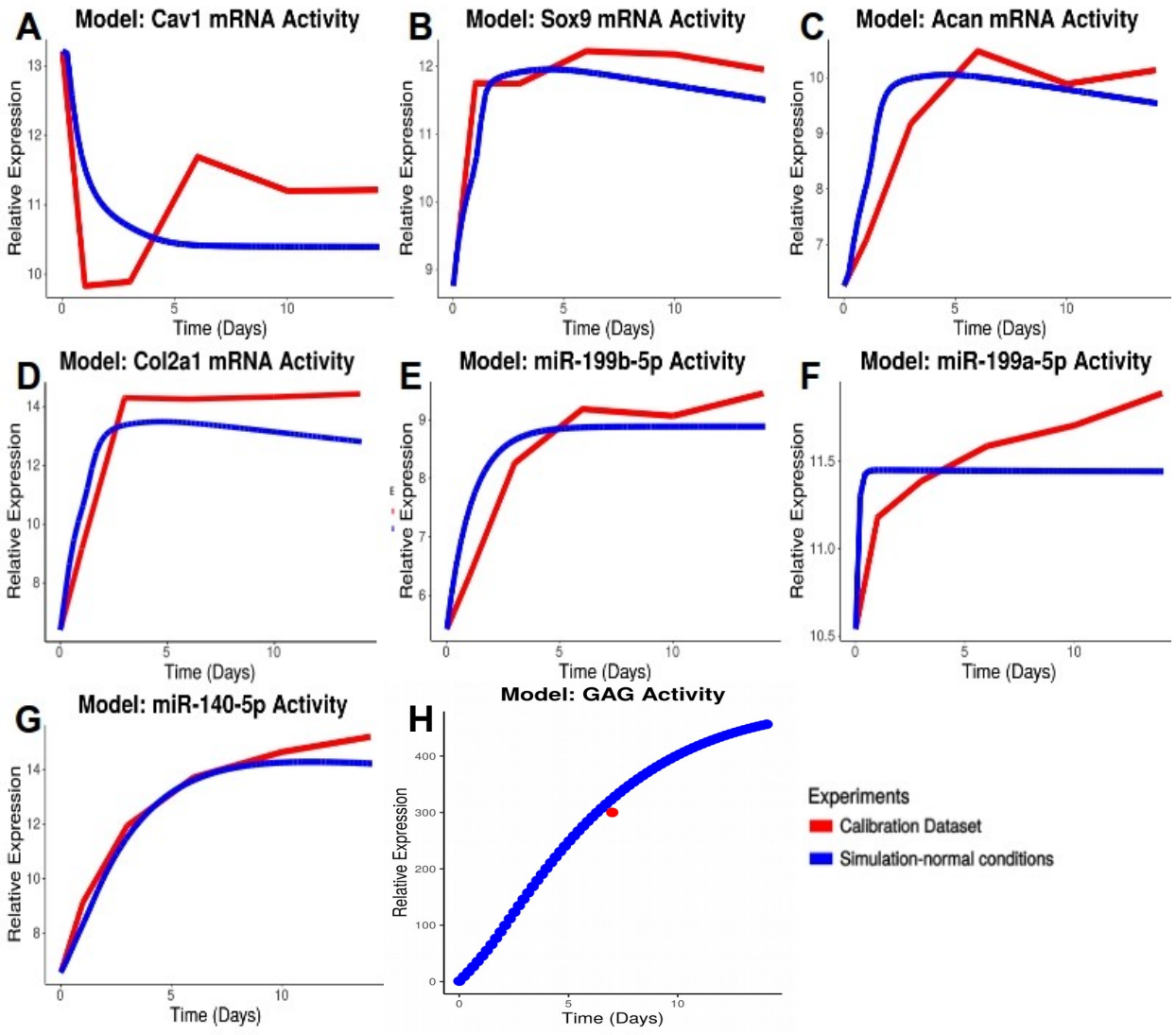


Figure 4.6: **Calibrated output from multi-miRNA chondrogenesis model.** Eight plots showing model objects from the microarray dataset (red) and simulations from the model (blue). Figure 4.6 H) shows a dotted blue line and one red dot (based on Figure 4.3D) because the experimental data was based on one time point.

Model Validation

Validation data from the qPCR results (Figures 4.3-4.4) was used as a barometer for how the model should be simulating under *miR-199b-5p* inhibition conditions. The model captures the inhibited behaviours well. Early *ACAN* and *COL2A1* simulations (day 1) do not reach the same nadir as the validation data. The inclusion of the 'OtherTargets' black box

helped to capture these dynamics. The 'OtherTargets' blackbox also supports the theory that *miR-199b-5p* targets other mRNAs which negatively regulate chondrogenesis. Also, the dynamics of the chondrogenesis targets and the 'OtherTargets' blackbox helped to predict the other *miR-199b-5p* targets would peak earlier than day three. This was because *CAV1* peaked at day 3, and in contrast the chondrogenesis biomarkers all share their nadir at day 1. Indicating other *miR-199b-5p* targets were earlier acting than *CAV1*. After day 7, the system was assumed to be going back to normal chondrogenic levels and the inhibition drug was predicted to have worn off around day 5.5. This was an assumption, however it was assumed to be after day 3 and before day 7, because by day 7 the chondrogenesis biomarkers begin to start matching their calibration levels. After day 7, all species were assumed to begin establishing a steady state. Another assumption was that the miRNAs were recycled in the system. It was possible to model the miRNA-mRNA interactions including several steps: miRNA-mRNA interaction complex formations, miRNA-mRNA binding and dissociation and miRNA degradation with the target mRNA. However, because our data was based over days, such fast reactions (which may span seconds-hours) would not be useful.

The following figures use the phrase 'Activity' to describe the behaviours of the mRNAs, because the transcript level measurements and simulations are being used as a proxy for gene activity. The other option was to use mRNAs as a proxy for proteins, but this would have been more complicated, as many of the proteins in this system undergo post-translational modifications e.g. phosphorylation. Using mRNAs as a proxy for overall gene activity was the simpler option.

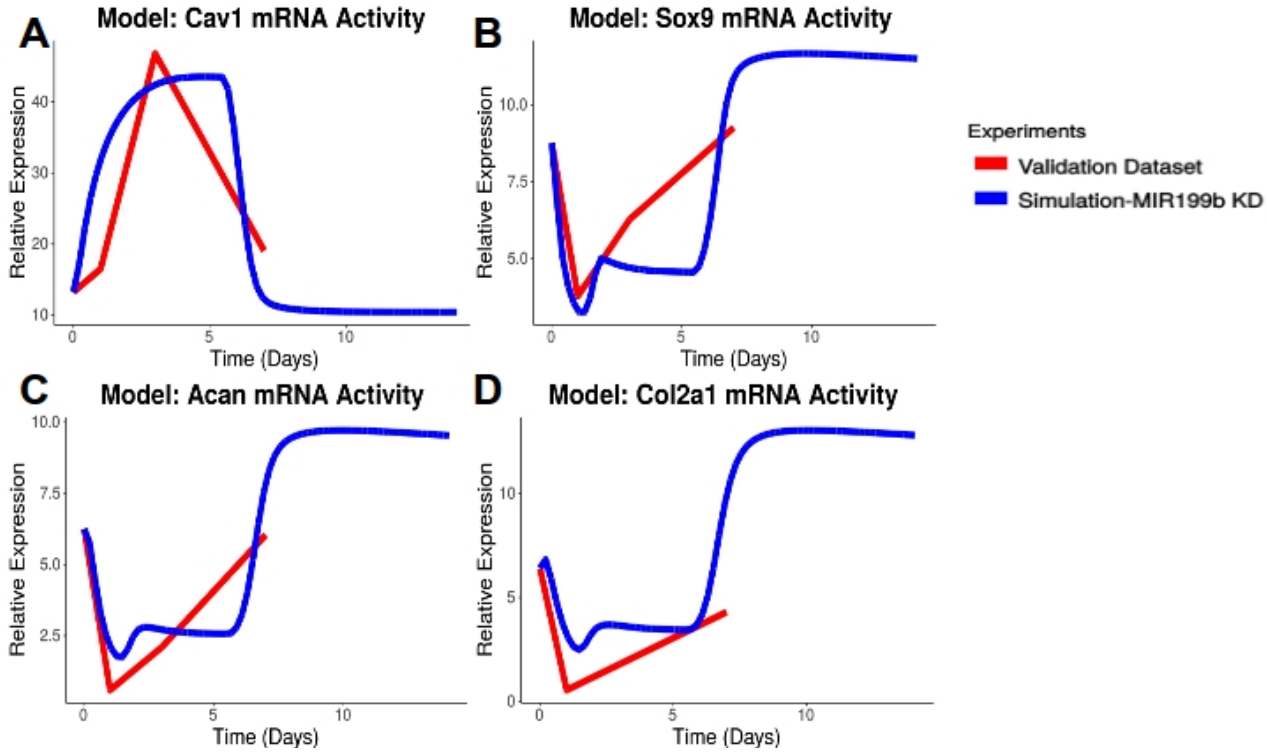


Figure 4.7: **Validation output from the multi-miRNA chondrogenesis model**. *CAV1*, *SOX9*, *ACAN* and *COL2A1* validation data (red) from Figure 4.3A-C and 4.4A contrasted against model simulation data during *miR-199b-5p* inhibition (blue). Simulations after day 7 are predicted based on the calibration data shown in Figure 4.6.

Predicting *miR-199a-5p* inhibition

miR-199a-5p inhibition simulations were predicted based on the assumption that *miR-199a-5p* would be performing the same function as *miR-199b-5p*, and also *miR-199a-5p* would have a slightly greater effect than *miR-199b-5p*. This assumption is based on there being a greater amount of *miR-199a-5p* found in the microarray dataset (Table 3.2). Based on this assumption, *miR-199a-5p* inhibition leads to a greater and earlier increase in *CAV1* and 'OtherTargets', when compared to *miR-199b-5p* inhibition. Furthermore, *SOX9*, *ACAN* and *COL2A1* simulations had greater nadirs during *miR-199a-5p* inhibition. During *miR-199a-5p* inhibition, the *ACAN* and *COL2A1* gene activity was closer to the validation data from figure 4.8. However, this also meant *SOX9* mRNA levels had a far greater down-regulation than what was found from the qRT-PCR experiments.

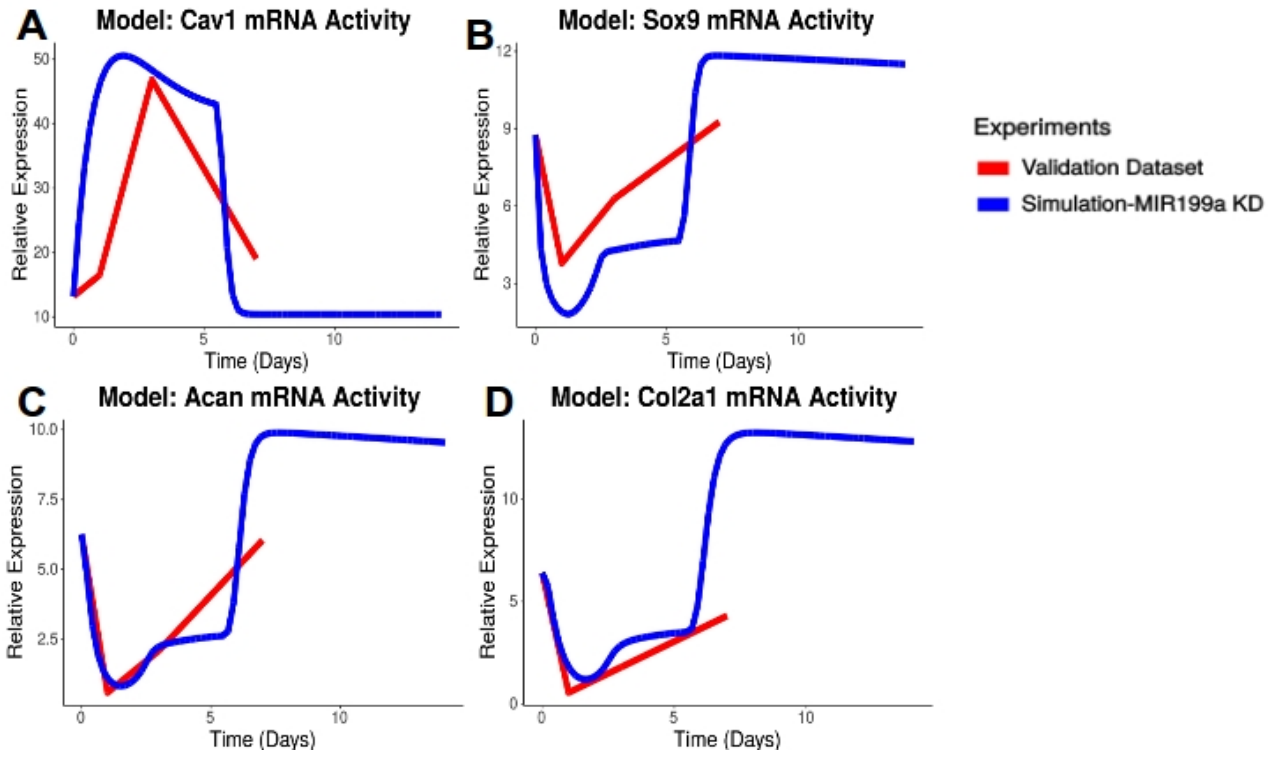


Figure 4.8: **Predicting affect of *miR-199a-5p* inhibition.** *CAV1*, *SOX9*, *ACAN* and *COL2A1* activity were simulated after *miR-199a-5p* inhibition. These simulations (blue) are contrasted against the validation data shown in Figure 4.3A-C and 4.4A (red). *miR-199a-5p* inhibition simulations were shown against *miR-199b-5p* inhibition results to show how the model predicts the effects of *miR-199a-5p* inhibition.

Further predictions from the model

miR-140-5p and GAG levels were predicted after *miR-199a-5p* and *miR-199b-5p* inhibition (Figure 4.9). For *miR-140-5p* and GAG levels, *miR-199a-5p* inhibition leads to a slightly greater drop. *miR-140-5p* was arguably the most important miRNA during chondrogenesis. It is promoted by SOX9 activity, thus when SOX9 decreases, *miR-140-5p* decreases. *miR-140-5p* activity will also promote GAG because *miR-140-5p* will degrade catabolic ECM genes such as *ADAMTS5*. GAG levels were measured to drop down by 60% on day 7 of chondrogenesis and this is simulated in Figure 4.9B. This simulation also allows us to predict GAG levels before and after day 7.

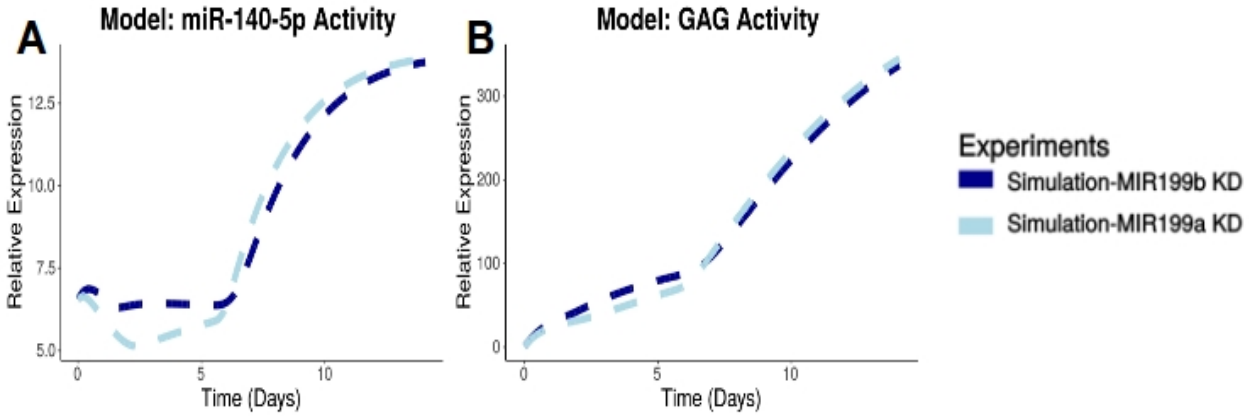


Figure 4.9: **Predicting *miR-140-5p* and GAG levels after *miR-199a/b-5p* inhibition.** Predictions of *miR-140-5p* and GAG activity during *miR-199b-5p* (dark blue) and *miR-199a-5p* (light blue) inhibition during the 14 day time course. Lines are dotted because *miR-140-5p* simulations are based only on calibration data and GAG simulation are only based on validation data during the 7 day time point (Figure 4.3D).

4.3 Methods

GRN construction

Upon further reading, GRNs were constructed in *CellDesigner* [318]. After several attempts the GRN shown in figure 4.2 was modelled in COPASI [119].

Kinetic Modelling

The model was calibrated using a mixture of parameter estimation and manual adjustments via sliders. Manual adjustments were needed due to the modelling software COPASI, not allowing multiple experiments (Caliberation and Validation) to work effectively simultaneously. Parameter estimation was mainly used for the model calibration. When using parameter estimation, particle swarm was used as the global algorithm and this was supplemented with Hookes-Jeeves, a local algorithm [119, 328, 329]. These algorithms aim to find the global minima for each species, by exploring the parameters [330]. The global algorithm functioned to find areas were the global minima may be, and the local algorithm used the end-point of the global algorithm to continue the analysis [331].

Most of the initial conditions and parameters were found using the microarray dataset from our collaborators [113]. Species which did not have any associated data had their initial conditions/ parameters assumed based on known behaviour from literature.

Data from qRT-PCR (Figure 4.3-4.4) had to be normalised to the microarray data to be used in the model. qRT-PCR results were calculated using the $\Delta\Delta CT$ formula [332]. qRT-PCR results was generated from three biological samples so the mean value and mean standard deviation were calculated for each sample. T-tests were performed between the controls and KDs to measure significance. To convert the qRT-PCR results to numbers which can be contrasted against the microarray results the following formula was used for each gene.

$$MEAN((Ig/Cg) * MRg) \tag{4.1}$$

- I = Inhibition value
- C = control value
- MR = microarray value
- g = given gene

Modelling was performed under the assumption that at day 0 the $MR = 1$. This way the calibration and validation data would have the same initial starting value (Table 4.4). This type of normalisation is necessary when parameterising a kinetic model with multiple datasets. T-tests were not carried out during day 0 because I worked under the assumption that at day 0, the inhibition and control data should be the same because an ODE based model requires the initial conditions of all species to be the same. If alternate initial conditions were used for the species, it would cause bugs when performing parameter estimation.

The model comprised of 16 species, 42 reactions and 4 events. Most species and reactions were within a single compartment (chondrocyte) and the model used d/mmol as its unit. GAG was stored in the ECM compartment. Both compartments = 1 and in a sense,

do not contribute to the model.

Table 4.4 shows some large disparities e.g. TGFB3 = 10000 and ACAN = 6.2571. Upstream species are kept at high concentrations to make the model initiate faster, and the alternative was to increase the size of the parameters. Note, naming conventions are dropped in the model, and for ease, species names were used instead.

Initial conditions

Species	Initial conditions
ACAN	6.2571
CAV1	13.2114
COL2A1	6.40998058
GAG	1
HP199b	0 (1 during event)
HP199a	0 (1 during event)
SOX9mRNA	8.76
SOX9Protein	1
MIR140_5p	6.55829
MIR199a_5p	10.5401
MIR199b_5p	5.42346
TGFB3	10000
SOX9PhosphoProtein	0
SRC	1000

Table 4.4: **Initial conditions from the multi-miRNA chondrogenesis model.**

4.3.1 ODEs

This is an ODE based kinetic model. Each species has specific inputs and outputs which control their behaviours over time. In each equation the term $ch = chondrocytecompartment$, and for GAG, $ecm = ECMcompartment$. All parameters have been rounded up to three decimal places. miR199b_amount and miR199a_amount are global quantities used to alter the miRNAs expression levels during events.

ACAN

$$\frac{d[ACAN].ch}{dt} = \begin{bmatrix} +ch.(\frac{100.[Sox9PhosphoProtein]}{1+\frac{1}{[Sox9PhosphoProtein]}}) \\ -ch.(4.263.[ACAN]) \end{bmatrix} \tag{4.2}$$

CAV1

$$\frac{d[CAV1].ch}{dt} = \begin{bmatrix} +ch.\frac{152.229.9625.57.[SRC]}{3.(0.416+[SRC])+9625.57.(0.1+[SRC])} \\ -ch.\frac{12.892.[CAV1]}{0.0972+[CAV1]+0.0971.\frac{[MIR199a_5p]}{0.1318}}.[MIR199a_5p] \\ -ch.\frac{18.746.[CAV1]}{0.0996+[CAV1]+0.0996.\frac{[MIR199b_5p]}{0.057}}.[MIR199b_5p] \\ -ch.(0.267528.[CAV1]) \end{bmatrix} \tag{4.3}$$

COL2A1

$$\frac{d[COL2A1].ch}{dt} = \begin{bmatrix} +ch.(\frac{94.624.[Sox9PhosphoProtein]}{1+\frac{1}{[Sox9PhosphoProtein]}}) \\ -ch.(3.005.[COL2A1]) \end{bmatrix} \tag{4.4}$$

GAG

$$\frac{d[GAG].ecm}{dt} = \begin{bmatrix} +ecm.(\frac{0.179.[COL2A1]}{1+89.324}/[COL2A1]) \\ +ecm.(\frac{4.70977.[ACAN]}{1+1}/[ACAN]) \\ +ecm.(\frac{3.97.[MIR140_5p]}{1+5}/[MIR140_5p]) \end{bmatrix} \tag{4.5}$$

MIR140_5p

$$\frac{d[MIR140_5p].ch}{dt} = \begin{bmatrix} +ch.(6.085.[Sox9PhosphoPortein]) \\ -ch.(0.376.[MIR140_5p]) \end{bmatrix} \tag{4.6}$$

MIR199b_5p

$$\frac{d[MIR199b_5p].ch}{dt} = \begin{bmatrix} +ch.(\frac{8.013.\frac{[TGFB3]}{0.0152]}}{1+\frac{[TGFB3]}{0.0152}}.miR199b\_amount) \\ -ch.(0.90175.[MIR199b_5p]) \\ -ch.(\frac{1.004.[miR-199a-5p]}{0.124}.[HP199a]) \end{bmatrix} \tag{4.7}$$

MIR199a_5p

$$\frac{d[MIR199a_5p].ch}{dt} = \begin{bmatrix} +ch.(\frac{100.41.\frac{[TGFB3]}{104.984}}{1+\frac{[TGFB3]}{104.984}}.miR199b\_amount) \\ -ch.(8.678.[MIR199a_5p]) \\ -ch.(\frac{1.009.[miR-199a-5p]}{0.0178}.[HP199a]) \end{bmatrix} \tag{4.8}$$

OtherTargets

$$\frac{d[OtherTargets].ch}{dt} = \begin{bmatrix} +ch.(\frac{1354.23.1554.29.[OtherTargetsRegulators]}{0.008.(0.198+[OtherTargetRegulators])+1554.29.(0.009+[OtherTargetsRegulators])}) \\ -ch.(\frac{100.728.[OtherTargets]}{0.1+[OtherTargets]+0.1.\frac{[MIR199a_5p]}{0.103}}) \\ -ch.(\frac{121.391.[OtherTargets]}{0.095+[OtherTargets]+0.095.\frac{[MIR199b_5p]}{0.105}}) \\ -ch.(0.097.[OtherTargets]) \end{bmatrix} \tag{4.9}$$

OtherTargetsRegulators

$$\frac{d[OtherTargetsRegulators].ch}{dt} = \begin{bmatrix} +ch.(0.099) \\ -ch.(8.652.[OtherTargetsRegulators]) \end{bmatrix} \tag{4.10}$$

SRC

$$\frac{d[SRC].ch}{dt} = \begin{bmatrix} +ch.(\frac{TGFB3}{100}.0.117) \\ -ch.(1.407.[SRC]) \end{bmatrix} \tag{4.11}$$

SOX9mRNA

$$\frac{d[SOX9mRNA].ch}{dt} = \begin{bmatrix} +ch.(604.499.[SOX9mRNA]) \\ +ch.(2.1577.[CAV1]) \\ +ch.(1.551.[TGFB3]) \\ +ch.(8.44.[MIR140_5p]) \\ -ch.([SOX9mRNA].11.921.[OtherTargets]) \\ -ch.([SOX9mRNA].59.112.[CAV1]) \end{bmatrix} \tag{4.12}$$

SOX9Protein

$$\frac{d[SOX9Protein].ch}{dt} = \begin{bmatrix} +ch.(6.085.[SOX9mRNA]) \\ -ch.(0.376.[MIR140_5p]) \end{bmatrix} \tag{4.13}$$

SOX9PhosphoProtein

$$\frac{d[SOX9PhosphoProtein].ch}{dt} = \begin{bmatrix} +ch.(6.085.[SOX9mRNA]) \\ -ch.(0.376.[MIR140_5p]) \end{bmatrix} \tag{4.14}$$

TGFB3

$$\frac{d[TGFB3].ch}{dt} = \begin{bmatrix} +ch.(1.848) \\ -ch.(0.00475.[TGFB3]) \end{bmatrix} \tag{4.15}$$

Events

Four events are used to simulate miR-199a-5p and miR-199b-5p inhibition. If triggered, HP199a_activity and HP199a_inactivity lead to HP199a = 1 until time reaches 5.5 days, at which point HP199a = 0. Likewise, if triggered, HP199b_activity and HP199b_inactivity lead to HP199b = 1 until time reaches 5.5 days, at which point HP199b = 0. HP199a and HP199b will reduce their target miRNA by 90% until day 5.5. The effectiveness of the *miR-199b-5p* inhibition was not measured, and this was a weakness in our model. I assume the inhibition of *miR-199b-5p* was effective, given the changes seen in chondrogenic biomarkers and *CAV1* levels (Figure 4.3-.4.4). HP199a and HP199b are global quantities which are fixed at 1. As global quantities they do not have inputs or outputs.

4.3.2 Functions

Several functions were used to capture the behaviour of the model species throughout chondrogenesis. Some functions were created for the model. All reactions were irreversible. Model specific terms are used here. A modifier/ M is a species which will affect

a substrate or parameter, but is not itself changed during the reaction. A substrate/ S is a species which is reduced during a reactions. A product/ P is a species which is produced from a reaction. A parameter (V, v, kn (n = #), Kms, Kac, Kms, ki) is a number which will act as a weight on a M or an S. Parameters are the targets to change when using sliders or parameter estimation to fit a kinetic model.

Constant Flux

$$v \tag{4.16}$$

v = parameter.

Constant Flux was used when importing TGFB3 and OtherTargetsRegulator into the model.

Declining input 1

$$\frac{V.k1.M}{Kms.(Kas+M)+k1.(Kac+M)} \tag{4.17}$$

V = parameter, k1 = parameter, M = modifier, Kms = parameter, Kas = parameter, Kac = parameter.

Declining input 1 was used when inputting OtherTargets(P) from OtherTargetsRegulator(M) and CAV1(P) from SRC(M). During these reactions the M will be declining to a low steady state at a fast rate, and the products (OtherTargets and CAV1) will decline at a slower rate until they reach a lower steady state.

HP modification 1

$$\frac{k1.S}{k2}.M \tag{4.18}$$

k1 = parameter, S = substrate, M = modifier, k2 = parameter.

HP modification 1 was used to set the amount of downregulation HP199a(M) and HP199b(M) have on their respective miRNAs(S). When events are not triggered, these functions do

nothing because M = 0. During events this function will be active because M = 1.

Fast input 1

$$\frac{V.M}{1 + \frac{k1}{M}} \tag{4.19}$$

V = parameter, M = modifier, k1 = parameter.

Fast input 1 was used when inputting ACAN(P) and COL2A1(P) from SOX9Phospho-Protein(M). This was also used when inputting GAG(P) from ACAN(M), COL2A1(M) and MIR140_5p(M). This was a fast reaction so the inputted species will follow the trend set by the modified.

Fast Input 2

$$k1.M \tag{4.20}$$

k1 = parameter, M = modifier.

Fast Input 2 was used when a product (MIR140_5p, SOX9mRNA, SOX9Protein) was required at speed comparable with mass action. However, unlike mass action this function uses an M instead of an S. Fast Input 2 worked faster than Fast Input 1 [118].

Mass Action

$$S.k1 \tag{4.21}$$

S = substrate, k1 = parameter.

Mass Action was the most common function this model, being used 16 times. It was used to output species, except for miRNA based output [118].

miRNA induced output 1

$$\frac{V.S}{Km + S + (Km.\frac{M}{K1})} \tag{4.22}$$

V = parameter, S = substrate, M = modifier, Km = parameter, Ki = parameter.

miRNA induced output 1 was used when outputting an S (CAV1 or OtherTargets) and a miRNA (MIR199b_5p or MIR199a_5p) was an M.

miRNA input 1

$$\frac{V.\frac{M}{shalve}}{1 + \frac{M}{shalve}}.GQ \tag{4.23}$$

V = parameter, M = modifier, shalve = parameter, GQ = parameter.

miRNA input 1 was used when inputting MIR199a_5p or MIR199b_5p. This function was a modified hill cooperative function as it used an M instead of an S [333]. This function also used GQ as a multiplier. During events the GQ parameter (HP199a or HP199b) was modulated to simulate *miR-199a-5p* or *miR-199b-5p* inhibition.

Mixed activation

$$\frac{V.S.M}{Kms.(Kas + M) + S.(Kac + M)} \tag{4.24}$$

V = parameter, S = substrate, M = modifier. Kms = parameter, Kac = parameter.

Mixed activation was used when SOX9Protein(S) became SOX9PhosphoProtein via CAV1(M) [334]. SOX9Protein was needed as the base of this reaction and CAV1 activity was needed as the proxy catalyst (ROCK1 is the true catalyst but was not present in the model

due to data limitations).

SOX9mRNA output from targets 1

$$(S.k1).M \tag{4.25}$$

S = substrate, k1 = parameter, M = modifier.

SOX9mRNA output from targets 1 was used when outputting SOX9mRNA(S) from MIR199a_5p or MIR199b_5p targets (CAV1 or OtherTargets)(M). This was a slower reaction as to not cause large levels of SOX9mRNA downregulation during non-event triggered simulations.

4.4 Summary

Chondrogenesis is a complex process which is regulated by miRNAs. Further study of this process using a systems and computational approach helped to identify *hsa-miR-199a-5p* and *hsa-miR-199b-5p* as potential pro-chondrogenic regulators and their novel target *CAV1* which is predicted to be an anti-chondrogenic gene. Experimental data provided some validation towards *miR-199a-5p-CAV1* and *miR-199b-5p-CAV1* as regulatory interactions during chondrogenesis. However, it seems clear the miRNAs have other targets which caused a larger and earlier anti-chondrogenic effect. Kinetic modelling allowed us to simulate the system with a ”black-box” which represented the alternative *miR-199a/b-5p* targets. This is the first validated multi-miRNA chondrogenesis model and allows us to simulate the effects of *miR-199b-5p* inhibition during chondrogenesis. The model can also predict the effects of *miR-199a-5p* inhibition which leads to a greater *CAV1* spike and thus a greater anti-chondrogenic effect. Furthermore the model can predict GAG and *miR-140-5p* levels upon *miR-199a-5p* or *miR-199b-5p* inhibition. The decrease seen in GAG levels indicated a loss of cartilage function.

CHAPTER 5

PREDISPOSITION MODEL FOR JUVENILE ONSET HUNTINGTON'S DISEASE

5.1 Background

ML is becoming an entirely unique field of computational biology because its usage has great prospects for better understanding biological processes and complex disorders such as neurological diseases [335]. ML can detect patterns from large datasets. The concept and full usage of this approach will not be discussed here because this would distract from my research. Instead these reviews highlight the power of ML in biological research [335, 336, 337].

Using ML to identify miRNAs of interest

Within the miRNA research community ML has mostly been used to predict miRNA-mRNA targets. This has proven very useful and ML based algorithms have enhanced the power of many prediction tools such as miRDB with uses an SVM (support vector machine)

model and TargetScan which uses a rules based model [79, 83]. However, there is also the potential of using ML to identify miRNAs as biomarkers. In subsection 1.1.5, I described how useful miRNAs could be as non-invasive biomarkers, so I will not labour this point here.

To showcase how ML can help to identify interesting miRNAs as biomarkers within complex diseases I searched for a large longitudinal miRNA-mRNA expression dataset in GEO. I found a large juvenile onset Huntington's disease (JHD) dataset from this repository - GSE65776. This contained data from mouse cortex, striatum and liver. The associated publications identified the striatum to be the most effected by HD, followed by the cortex and no significant effects were seen in liver samples. I downloaded the cortex dataset because it was the most complete [114, 115]. To identify a suitable ML strategy appropriate for the dataset, it was analysed by differential expression using *DESeq2* and then analysed with *TimiRGeN*.

5.1.1 Background biology of HD and juvenile onset

HD is a chronic neurodegenerative disorder characterised by the progressive loss of neurons. Patients with HD show signs of loss of motor skills and memory, along with reduced capability of maintaining normal living conditions without aid [338]. Brain scans of HD patients reveal a loss of neurons in the striatum and cortex. Cortical thinning is thought to be an earlier pathological event [339].

HD is a rare condition, affecting 10.6-13.7 individuals per 100,000 in Western populations. Prevalence of the disease is significantly lower in east Asian and black populations [340]. HD is caused by a CAG codon expansion in exon 1 of the *HTT* gene. HD is an autosomal dominant disease so only one mutant allele is enough to lead to disease phenotypes. There is currently no cure and patient mortality is 100%. In terms of molecular changes which lead to this disease; a CAG expansion leads the *HTT* gene expressing a mutant *HTT* (mHTT) mRNA isoform with an extended 3ÚTR. The mHTT mRNA isoform will transcribe a truncated mHTT protein with a long glutamine (poly-Q) chain. mHTT has been recorded to display alternate phenotypes to wild type HTT (wtHTT) . The specific mutation

is termed a poly-Q expansion and the specific mutation of an individual can be measured by genotyping the number Q repeats on the *HTT* gene. mHTT protein promotes neurotoxicity, and mortality from HD results from decay of white and grey matter [338].

CAG repeat numbers are assessed during genetic screens to identify the likelihood of the disease being present in an individual. Often individuals with 20 of fewer CAG repeats will have no onset, though some papers have reported the normal range to extend to 27 CAG repeats [114, 341]. An intermediate range of penetrance of 27-35 CAG repeats has largely been accepted by the literature, and full HD penetrance in individuals with $>$39 CAG repeats [341, 342]. Phenotypes usually develop during the middle aged years (30 to 50 years). However, another class of CAG repeats can lead to juvenile onset HD ($<$25 years old). Here the CAG repeats are $>$60, though there is some debate on the threshold [343]. As with normal onset HD, JHD is more prevalent in Caucasian groups [344]. Specific reports identify 12.3 per 10,0000 individuals in the UK and 9.3 per 10,0000 in Germany [345, 346]. Interestingly, CAG repeat length only accounts as one of several factors when it comes to the severity of HD. Twin studies found altered severity, indication epigenetic or environmental factors also play a role in HD onset and severity [347].

Molecular changes from mHTT expression

The wtHTT protein is ubiquitously expressed, and in striatal and cortical neurons it can be seen in the cytoplasm [348]. It is a $\tilde{3}$50 kDa sized protein with many HEAT binding regions. These motifs are vital for protein-protein interactions. HEAT motifs are used in scaffolding roles [349]. The wtHTT protein also has a nuclear export sequence and the poly-Q expansion mutation often causes impaired function of this sequence. mHTT proteins can become unable to leave the nucleus and aggregates inside the nucleus [350]. As a scaffolding protein, HTT binds with B-tubulin, microtubules, dynein/ dynactin. The ability to bind with B-tubulin and microtubules also means HTT has extensive functions as a cellular trafficking protein [351, 352]. The HTT protein also has transcriptomic regulatory effects. For example, HTT binds to and sequesters the transcription factor repressor element-1 transcription factor (REST), which is a negative regulator of brain-derived neurotrophic factor *BNDF*. BDNF promotes neuron survival. mHTT has reduced capacity to sequester

REST, and therefore, *BDNF* downregulation increases in HD patients [353, 354].

mHTT protein is thought to contribute to a diseased state via several altered mechanisms. This includes: the poly-Q expansion leading to protein cleavages which aggregate and may sequester essential proteins leading to toxicity, altered transcriptional expression of genes like *BDNF*, altered protein trafficking, and also altered mRNA and miRNA expression [355].

miRNA interest in HD

Expression studies of HD have found altered miRNA expression. This is not surprising because of the range of functions the HTT protein has. Also, mHTT has been shown to alter transcription factor activity e.g. REST and BDNF, and the altered activity of these transcription factors could lead to altered miRNA expression in HD patients. Several miRNAs have been detected as SDEGs from bioinformatics studies. For example, *miR-9* and miR-9* have been found to be downregulated in early HD cortical samples and may correlate negatively with HD severity [356]. *miR-124* has been seen upregulated and downregulated in HD patients, making its function in HD unclear. It should also be noted that *miR-9, miR-9** and *miR-124* target *REST* and *coREST* (RCOR1) mRNAs for downregulation [356, 357]. RCOR1 is a protein partner of REST, and also functions as a repressor protein and a mediator of REST function [358].

5.1.2 Preliminary data analysis

168 individual mice were sacrificed and their cortex were harvested. RNAseq and miRNAseq was performed on each cortex, thus 336 sequencing experiments were carried out [114, 115]. The mouse samples could be broken down into gender (male, female), Q mutation (WT, Q20(WT), Q80, Q92, Q111, Q140, Q175) and age (2M (month), 6M, 10M) of sacrifice (Table 5.1). The Q20 mutated mice were positive controls and can be used to check the viability of the WT, as there should be few-no SDEGs found between WT and Q20.

Some samples were removed because they were identified as outliers based on PCA plots (not shown). Also, it should be noted that the 6M data was distinct from the 2M and 10M data because it had a far greater number of differentially expressed genes than expected. The authors of the publication did not report an abnormally high number of SDEGs from their cortex samples, but did report this from the liver samples, and (to a lesser extent), their striatum samples. I believe the 6M data may have experienced a batch effect and the authors of the publication used *ComBat*, a batch correction method from the *sva Bioconductor* tool, to re-center the data [114, 359]. Though the authors state they only used *ComBat* to reduce noise from gender differences, but as the code is not available online, I could not check how batch correction was performed. 6M data was kept in the analyses as this assessment may have just been my speculation and also *TimiRGeN* analysis required at least three time points to function. It is my intention to repeat the analysis presented in this chapter without the 6M data.

Q-mutation	Gender	2 month		6 month		10 month	
		mRNA	miR	mRNA	miR	mRNA	miR
WT(Q20)	M	4	4	4	4	4	4
Q20	M	4	4	3	4	4	4
Q80	M	5	5	4	4	4	4
Q92	M	4	4	4	4	4	4
Q111	M	4	4	4	4	4	4
Q140	M	4	4	4	4	4	4
Q175	M	4	4	4	4	4	4
WT(Q20)	F	4	4	4	4	4	4
Q20	F	4	4	4	4	5	4
Q80	F	3	3	4	4	3	4
Q92	F	4	4	4	4	4	4
Q111	F	4	4	4	4	4	4
Q140	F	4	4	4	4	4	4
Q175	F	4	4	4	4	4	4

Table 5.1: **Spread of HD expression dataset.** The samples can be divided by age, gender and Q mutation for miRNA and mRNA samples. Most conditions had four samples, and a few had three or five.

Circulating miRNAs are potential biomarkers for JHD

The fact that miRNAs have been detected as changing in the cortex of HD patients opens the possibility of finding circulating overexpressed miRNAs in blood plasma or cerebrospinal fluid (CSF) which could act as biomarkers for HD [360]. This would be especially interesting if these miRNAs could be used to detect predisposition to JHD. Circulating miRNAs have been investigated in the blood plasma and CSF of HD patients [361, 362]. *hsa-miR-34b* has been identified as a promising plasma stable biomarker [361]. However, as in the case of the renal diseases discussed in subsection 1.1.5, accurate diagnostics require multiple circulating miRNAs. This is important, as other neurodegenerative diseases such as Parkinson's and Alzheimer's may also lead to over-expression of circulating miRNAs

in blood plasma and the CSF [360, 363]. The dataset analysed in this project is from the cortex, so any miRNAs found here may only be cortex specific. It would also be interesting to see miRNAs that are changing in the striatum.

5.2 Results

To establish a successful means of attempt at applying ML to identify biomarkers for JHD, I first explored the data to find a suitable and straightforward question which could be answered by ML. This data centric A.I approach is promoted by Stanford Lecturer and ML practitioner Andrew Ng [364]. *DESeq2* and *TimiRGeN* were utilised to identify a suitable narrative for ML [117, 145]. The analysis found age to be the best variable to assess and also found gender and severity of Qvalue mutation to be less important variables. From this I decided to use ML to classify samples as WT or HD. To make this an early predisposition project, the 6M and 10M data were used for model training and the 2M data was used as an independent test dataset. Feature engineering and feature selection methods were also used to find suitable genes to train on. Multiple classifiers were used to train and test different models. Logistic regression performed the best, so was used to create a confusion matrix and ROC/AUC curve to show how well the model performed.

5.2.1 Data exploration with DE and *TimiRGeN*

Age (2M, 6M, 10M), gender (M, F) and Qvalue (WT, Q20, Q80, Q92, Q111, Q140, Q175) mutation were the three variables in this dataset. Different samples in this section are denoted using this formula xM_yG_zQ / xM_yG_WT, where x represents age, y represents gender and z represents QValue mutation. miRNA and mRNA data were explored initially with DE and then in some cases with *TimiRGeN*. Outliers removed from the analysis were removed from both the miRNA and mRNA datasets, this was to keep a consistent number of samples during ML, as blank values/ zeros are not ideal for training.

Qvalues

With only Qvalues changing at each instance, each xM_yG_zQ was contrasted with their corresponding xM_yG_WT with pairwise DE. For miRNA and mRNA data alike, very few

SDEGs were found from the 2M and 10M analyses. miRNA data found between 0-3 SDEGs and mRNA data found between 0-16 SDEGs. The 6M analyses often had over 3000 SDEGs ($<$0.05 adjusted Pvalue value). Even $6M\_M\_Q20/6M\_M\_WT$ and $6M\_F\_Q20/6M\_F\_WT$ respectively found 1597 and 2542 SDEGs. Q20 mice were a control and genetically the same as the WT mice, so there was clearly an issue with the 6M data. This analysis was not carried forward to *TimiRGeN* because only a few SDEGs were found in the 2M and 10M analysis.

Gender

To explore differences in JDH between genders a similar analysis was performed as above. With gender being the only changing variable. Each xM_M_zQ was contrasted with their corresponding xM_F_zQ for pairwise DE. This analysis was performed by Bethany Harley, a project student who I co-supervised. Based on the previous analysis, the WT and Q20 data were treated as the same sample, because no SDEGs were found between the WT and Q20 samples, which was expected (except in the 6M data) [114, 115]. Like with the Qvalue analysis, few miRNAs and mRNAs were found to be SDEGs (except in the 6M data). Results from DE were taken forward for *TimiRGeN* analysis. Few pathways were found to be enriched. The only trend seen was an enrichment in cholesterol synthesis related pathways, though this was only seen in a few Qvalues. Cholesterol synthesis is important in production of the myelin sheaths which are known to degrade in several neurodegenerative diseases and the loss of myelin sheaths along axons contributes to nuerotoxicity [365, 366].

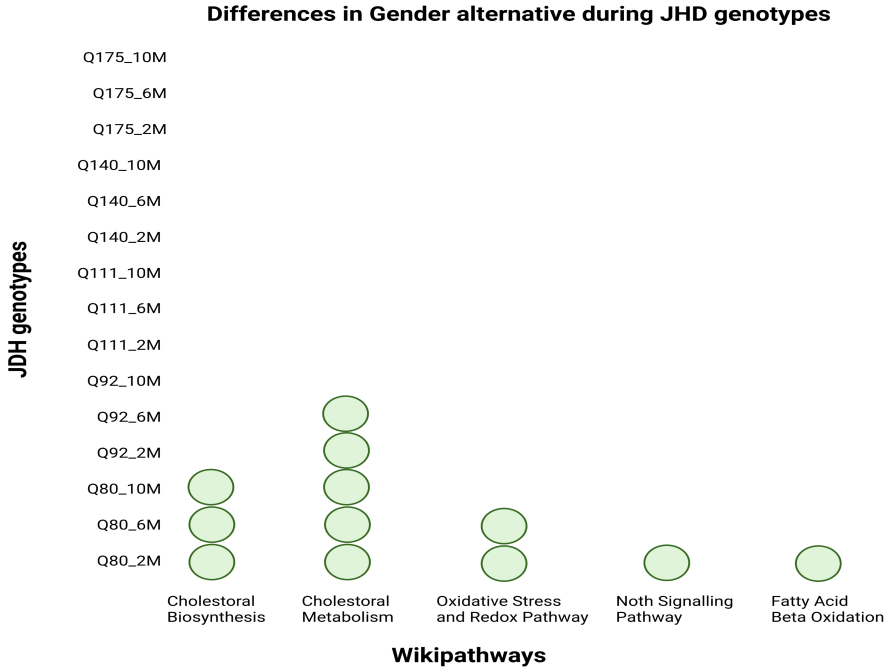


Figure 5.1: **Pathways found from analysing gender based SDEGs with *TimiRGeN*.** SDEGs from gender based DE are analysed with *TimiRGeN* and the associated pathways are shown here. Pathways include: Cholesterol Biosynthesis, Cholesterol Metabolism, Oxidative Stress and Redox Pathways, Notch Signalling Pathway and Fatty Acid Beta Oxidation. Enriched pathways were found towards the lower scale of Qvalue mutations. Data for this image was taken from Bethany Harley's UG project.

Age

Qvalues and gender were not found to be very interesting variables. So when exploring age as a variable of interest, males and females were combined into a single class, and Qvalues were sorted into HD or WT conditions. Q20 and WT samples were classed as WT. Therefore, now each class could be summerised by xM_HD or xM_WT. 2M, 6M and 10M SDEGs respectively numbers as 22, 127, 83 for mRNAs and 1, 105, 6 for miRNAs. This data was not further analysed with *TimiRGeN* because, as the previous analysis with the gender based differences showed, not enough SDEGs are found for proper analysis with pathway enrichment.

Early detection of JHD

From the data analysis it was clear that age was the most important factor. Gender and Qvalues most likely contributed small amounts of variance between samples. Combining genders and classifying Qvalues into WT (WT and Q20) and HD (Q80, Q92, Q111, Q140, Q175) lead to far more repeats in the following classes: 2M_WT (15), 6M_WT (15), 10M_WT (15), 2M_HD (39), 6M_HD (39), 10M_HD (40); which was useful for model training. One interesting question would be to see if SDEGs (P adjusted value is lower than 0.05) found in both 6M and 10M mice could be used to identify if 2M mice were predisposed for HD or were WT. This is because the 6M and 10M JHD onset mice should show signs of HD and the 2M mice would not be showing symptoms at the time of sacrifice. Furthermore, using genes from both 6M and 10M data increases confidence of the genes being put forward for ML analysis being biologically important to HD, because again, I am skeptical on the validity of the 6M data.

31 SDEGs from both 6M and 10M data were identified from the mRNA and miRNA data: *Anln, Asrgl1, Atraid, Bhlhe41, Car2, Cd209c, Chdh, Cldn14, Fads1, Gdf10, Gm21168, Gm5067, Gm6089, Hey1, Il17rb, Il33, Nrf1, Nudt4, P4ha3, Plpp3, Plxnb3, Rrs1, Slc45a3 ,Teddm2, Tlcd1, Tmc3, Tmem40, Wnt10b, mmu-miR-135b-5p, mmu-miR-212-3p, mmu-miR-221-3p*. These genes were used as the independent features for training models. Normalised expression values from these genes were extracted from the 2M data (test set) and the 6M and 10M data (train set). The 31 genes were not checked if they were SDEGs in the 2M data as this would lead to bias. The dependent feature was named Sample and this contained the string 'HD' or 'WT' for each sample. *scikit-learn*/*Python3* were used for ML analysis [367, 368].

5.2.2 Early detection of JHD using ML

Remove highly correlating features

Data exploration is required before testing ML models because removing highly correlated features from the training data reduces risk of over-fitting. Removing lowly correlated features could also be helpful, but these were not found because of the feature engineering approach used DE. A Spearman correlation heatmap was created from the training data

(Figure 5.2). If two or more features correlated at a rate of $=< 0.8$ or $>= -0.8$, only one of the features was kept. *Il33, Gm6089, mmu-miR-212-3p* and *Plpp3* were removed from the training and test datasets, thus 27 genes were used for classifier testing.

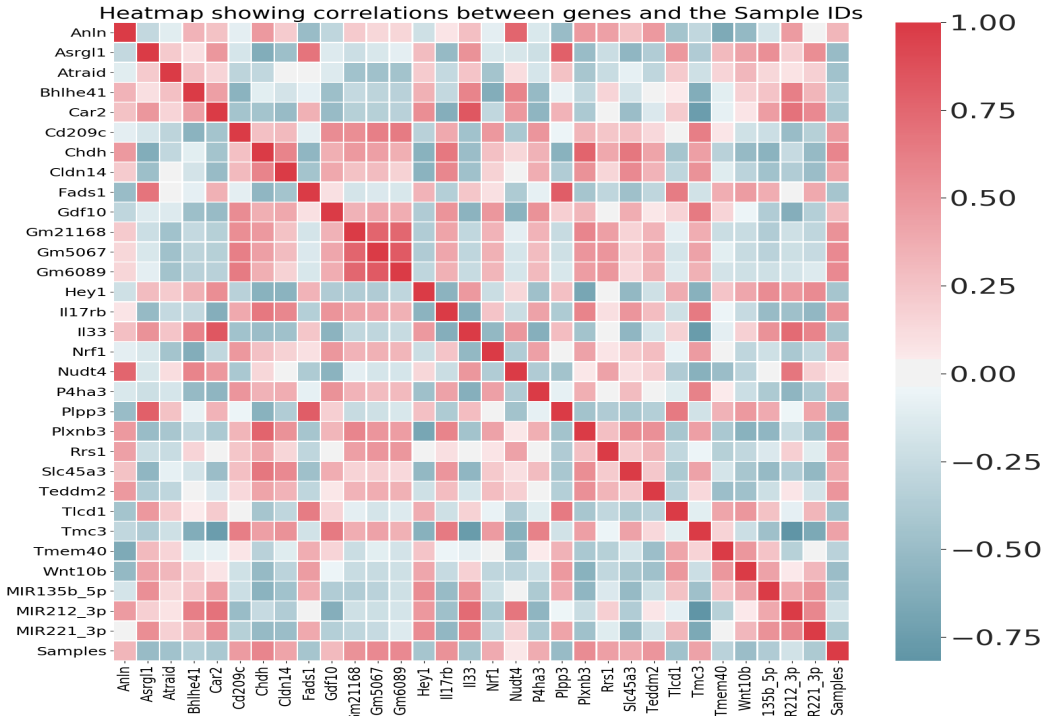


Figure 5.2: **Heatmap showing correlations between the features and the Samples.** Samples (HD or WT) were correlated with each feature. The 31 features were also correlated with each other to determine if features had similar patterns. Blue shaded boxes are negatively correlated and red shaded boxes are positively correlated. Spearman was used as the method.

SMOTE

Training and testing datasets respectively contained 78 HD and 30 WT samples and 39 HD and 15 WT samples. Many classifiers will bias towards the larger condition. SMOTE from the *imblearn* package was used to generate synthetic data based on the WT samples [369]. However, if cross-validation was used on synthetic data, it would no longer serve as a real world example, and lead to data leakage. For this reason SMOTE was used inside a cross-validation pipeline called *SKFold*. *MinMax* scaling was also added to this pipeline,

but scaling was performed before SMOTE because scaling synthetic data may have also lead to bias [370].

Comparing classifiers

The SMOTE/scaling pipeline was applied to test several popular classifiers. 5 shuffled cross-validations were performed to train each classifier. Mean results were calculated to compare how well each classifier performed [368]. Fifteen different classifiers were contrasted with the the training and validation data. The classifiers used were:

1. KNeighborsClassifier(n_neighbors=5, algorithm='auto', weights='distance')
2. SVC(kernel="linear", C=0.5)
3. SVC(kernel="poly", degree=3, C=0.025)
4. SVC(kernel="rbf", C=0.025, gamma=2)
5. GaussianProcessClassifier(1.0 * RBF(1.0))
6. GradientBoostingClassifier(n_estimators=100, learning_rate=0.001)
7. DecisionTreeClassifier(max_depth=3)
8. ExtraTreesClassifier(n_estimators=10, min_samples_split=5)
9. RandomForestClassifier(max_depth=3, n_estimators=100)
10. MLPClassifier(alpha=0.001, max_iter=10000, solver='sgd')
11. AdaBoostClassifier(n_estimators=100, learning_rate=5)
12. GaussianNB()
13. QuadraticDiscriminantAnalysis()
14. SGDClassifier(loss="hinge", penalty="l2")
15. LogisticRegression()

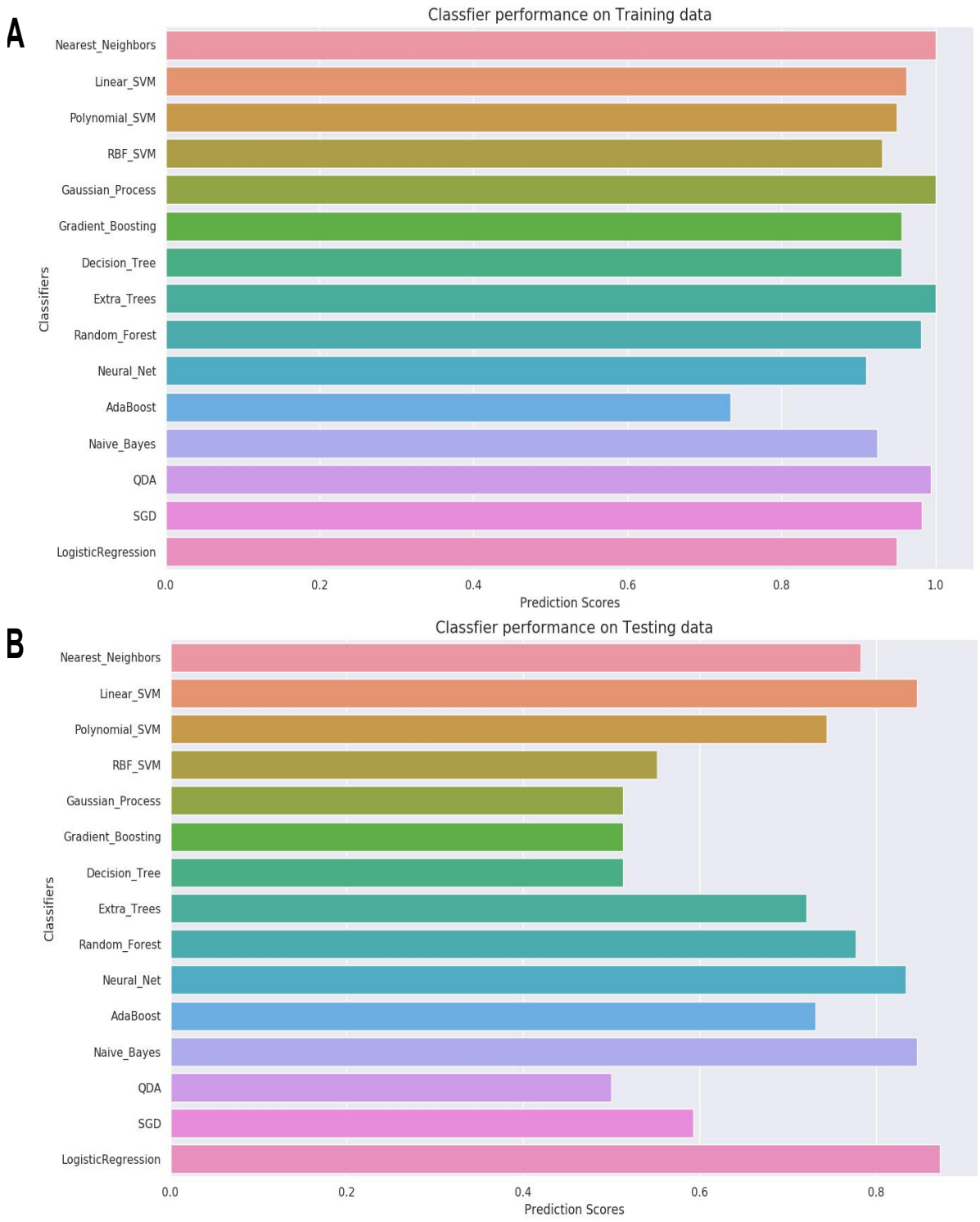


Figure 5.3: **Classier performance when applied to training and validation data.** Mean results from fifteen classifiers on the **A)** training data (6M and 10M) and **B)** the testing data (2M).

Results for the performance of each classifier is shown in Figure 5.3. Apart from AdaBoost, all of the classifiers trained above 90%, some even training at 100%, including: Nearest_Neighbors, Gaussian_Process and Extra_Trees. However, few managed to predict the correct samples in the test dataset above 80%, and this included Linear_SVM, Neural_Net, Naive_Bayes and LogisticRegression, the lattermost achieved the highest prediction at 87%. This clearly shows our models tend to under-fit.

Logistic Regression performance

Scaling, SMOTE and model training with Logistic Regression were wrapped around randomly shuffled stratified 5-fold cross-validation. Mean statistics were calculated to construct a confusion matrix and a ROC/AUC plot (Figure 5.4).

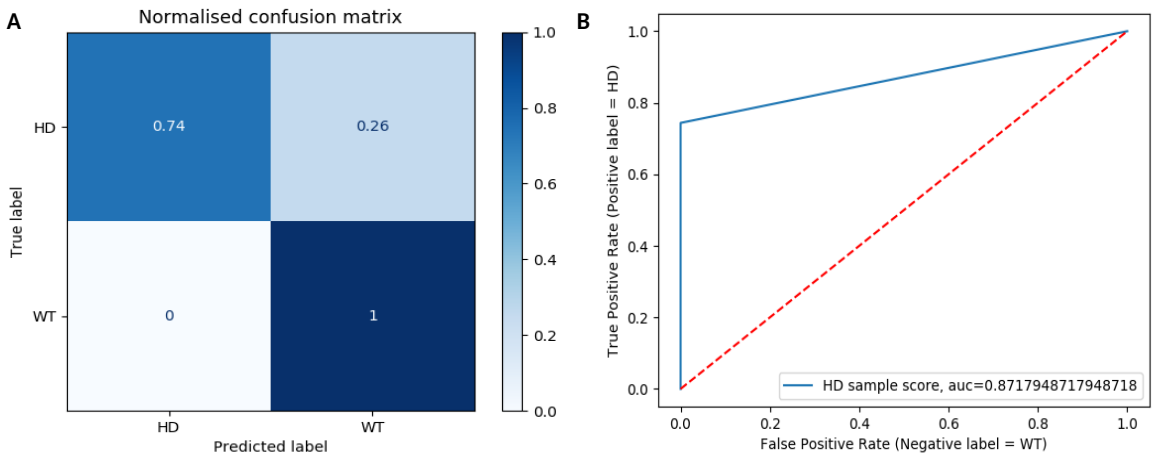


Figure 5.4: **Results from from Logistic Regression model. A)** Normalised confusion matrix contrasting true HD and WT labels. **B)** ROC/AUC curve comparing the true positive rate and the false positive rate.

The results above answer the question: how well can we determine the HD samples from the WT samples. The confusion matrix shows 100% accuracy when predicting WT samples, however it struggles to distinguish some HD samples. 74% of the HD samples were predicted correctly, while 26% were predicted incorrectly. The ROC/AUC plot (Figure 5.4B) shows a 87.1% accuracy when trying to detect label HD samples.

In this model, HD samples correctly labelled as HD are true positives, HD samples incorrectly labelled as WT are false positives, WT samples correctly labelled as WT are true negatives and WT samples incorrectly labelled as HD are false negatives. The Y axis of the ROC/AUC curve is the true positive rate which is calculated by the following equation.

$$\frac{TruePositives}{TruePositives + FalseNegatives} \tag{5.1}$$

The x-axis is calculated by $1 - Specificity$ which is equal to the false positive rate.

$$\frac{FalsePositives}{FalsePositives + TrueNegatives} \tag{5.2}$$

Overall this means the ML model characterised HD samples well, however the model has room for improvement. As it stands, this model would not be suitable for publication as mischaracterisation of a serious illness by $< 13\%$ (according to the ROC curve) means the model may not be useful for further research.

5.3 Methods

Data processing

miRNA and mRNA data from GSE65776 were downloaded using *SRA-toolkit* [188, 368]. Mus_musculus.GRCm38.cdna.all.fa was used by *Bowtie* to create index files for the mouse transcriptome. *miRDeep2* was then used to quantify mature mmu-miR strands from the miRNA data and the mature reads could be read into *R* [194]. *FASTQC* was used to to identify sequencing quality [189]. No adapter sequences were identified. *Salmon* was used to process the mRNA data and the sequence quality was acceptable [193]. In *R*, *tximport* was used to import the mRNA data [371]. Once in *R*, *DESeq2* was used to perform pairwise DE on the miRNA and mRNA data separately, and this identified SDEGs

based on age, gender and Qvalue [145]. The combined mode of *TimiRGeN* was used when looking for time course patterns from SDEGs found from gender based DE analysis [117]. Based on the data exploration, WT and Q20 samples were classed as WT samples, and all other samples were classed as HD samples. 31 SDEGs were identified from DE between WT and HD samples in both the 6M and 10M data. Normalised data for 31 genes are extracted from the 2M data to create a test set and the 6M and 10M data to create a train set.

Machine Learning with scikit-learn

Using *Python 3.6* the *scikit-learn (sklearn)* library, classification of HD and WT samples was performed [367, 368]. SDEGs from the 6M and 10M data were used for training and these same genes were used for testing. The training data was checked using a Spearman correlation based heatmap. Four features/ genes were removed because they had high correlation with other features/genes. 5-fold Stratified cross-validation was performed to split the training data into five parts, four of the parts would be used for training and one part would be used for testing. This would be repeated five times, and each time random data points were used. In each fold, data was scaled and then synthetic data points were made for the WT samples using SMOTE from the *imblearn* package [369]. This leads to an equal number of HD and WT samples, reducing classifier bias towards the more popular condition. Fifteen classifiers were contrasted and Logistic Regression performed the best so a normalised confusion matrix and a ROC/AUC curve was made with this classifier.

5.4 Summary

During this project I carried our preliminary work towards identifying potential miRNA biomarkers for JHD. ML is a very versatile and powerful tool. I created a model using Logistic regression which identified 2M HD samples at a perfect rate. However, this model only identified 74% of 2M HD samples as HD. This may be a product of the 6M data being effected by a batch effect. I intend to repeat this work without the 6M data and contrast the differences seen. Longitudinal miRNA-mRNA expression datasets as large, or larger than

the JHD data investigated here may be generated more often in the future. ML techniques should be theoretically and practically understood to investigate these datasets. Here I created a ML project to investigate how to utilise ML techniques on Longitudinal miRNA-mRNA expression datasets.

CHAPTER 6

GENERAL DISCUSSION

As stressed several times in this thesis, analysis of longitudinal multiomic datasets is difficult because many current tools cannot handle longitudinal datasets. Thus, novel methods of analysing such complex expression datasets must be established for the research community. During this PhD, I have utilised three computational techniques to investigate longitudinal miRNA-mRNA expression datasets, which are: the development and application of the *TimiRGeN R* package, creation of a multi-miRNA kinetic model which I used to explore how *miR-199b-5p* may regulate chondrogenesis and also to predict how *miR-199a-5p* regulated chondrogenesis, and I employed ML techniques to identify features the for predisposition of JHD [117]. These methods are explored through this thesis and I have shown their use with a range of complex time series datasets. This includes data from kidney injury, chondrogenesis, breast cancer and HD research [109, 110, 111, 112, 113]. In this discussion chapter I will outline the positive and negative aspects of each of the results chapters, describe further work not included in this thesis and highlight how each chapter contributes to biological research.

6.1 Ch2 - TimiRGeN R package

The *TimiRGeN R* package was developed as an effort to identify *miRNA-mRNA* interactions of interest from large multiomic datasets that have been measured along a time

course. It was also an effort to further extract information after DE analysis. DE is a useful tool, however at times it serves as a ranking metric. Biology is fluid and dynamic, and ranking metrics like DE do not grasp complex biological activity. Thus, I created *TimiRGeN* to act as a downstream analysis tool which can be used to supplement DE analysis of longitudinal miRNA-mRNA datasets [117]. It became the basis of an original paper in *Bioinformatics* and was accepted onto *Bioconductor*. The tool fills an analytical niche between DE analysis and hypothesis generation when analysing longitudinal miRNA-mRNA datasets. Furthermore the tool can accommodate a variety of inputs which was rare among miRNA based *Bioconductor* tools (see subsection 2.1). *TimiRGeN* can be used: after non-pairwise or pairwise DE techniques, for microarray or RNAseq datasets, for only miRNA or mRNA data, on static datasets, on a range of higher eukaryotic species (cow, dog, human, mouse and rat have been tested) and for analysis of miRNA-mRNA data combined or separately. This provides users with flexibility when designing their bioinformatic pipelines. This is a tool which I would like to invest time in maintain (if future employment allows) and there are several areas in which the tool can be improved.

Wider range of species

One of the most common questions asked about the package during conferences (see Appendix B) was if tool usage would be expanded to cater for a wider range of species. Common requests included plant species like *Arabidopsis thaliana* and lesser eukaryotes such as *Saccharomyces cerevisiae*. I plan to see how feasible this level of inclusion would be. Though this may not be a straight-forward task as adding plant/ micro-organism annotations. miRNAs function differently in plants e.g. miRNAs have a greater catalytic function in plants as the plant miRISC complex's are known to directly cleave their target mRNAs, in contrast mammalian miRISC complex's are more commonly known to use decapping to trigger mRNA decay (see subsection 1.1.2). Plant and mammalian miRNAs also have differences in biogenesis procedures, shape of their pre-miRNAs and plant miRNAs are known to bind with target sites in the open reading frame of plant mRNAs, while mammalian miRNAs almost exclusively target regions in the 3' UTR of target mRNAs [372].

***Cytoscape* focused updates**

Another comment I had during a conference was that I am not making the most of the potential *Cytoscape* has to offer for GRN development. This is true as I focused on *R* - *PathVisio* cross-platform access for GRN design, and I hope to explore further potential cross-talk between *R* - *Cytoscape*. *Cytoscape* is a very widely used tool and increasing attention here may also increase interest in *TimiRGeN* [121].

Improve automation to PathVisio

Currently *R* - *PathVisio* automation is limited. I have attempted using the *Rpathvisio* API with little success [373]. Once *TimiRGeN* is more established, I plan on contacting the maintainers of *PathVisio* and ask if they plan on developing automation between *R* and *PathVisio*. I have asked questions on this topic on the *PathVisio*/ Wikipathway forum, but I did not get any response.

Continually check for bugs and update the package appropriately

In the *Bioconductor* repository, I have included a link to an issues page which is associated to the *TimiRGeN* github site, for users to report bugs or issues. I plan to continue monitoring the package for bugs and to work with the community of users to make sure this package is usable for many years to come. Every 6-8 months *Bioconductor* enters a new development cycle, which is compatible with the newest developmental version of *R*. This can lead to code changes and thus lead to bugs to fix.

The *TimiRGeN* package is the cornerstone of this PhD and analysis with this tool has guided work in the other chapters too. Though, while this tool has great capacity to guide *in silico* work, and like most hypothesis generation tools, it works best in tangent with validatory data.

I also want to mention, methods subsection 2.3.1. Package creation was difficult, and unfortunately no one in my immediate group had experience with developing in *Bioconductor* before. Sources on *Bioconductor* specific package development were very limited. Few

pages on their website were useful. Comments from stack overflow and the other forums were also unhelpful because *Bioconductor* had specific requirements e.g. the vignettes should not be built in the submitted package, and the an extdata file should be in the inst folder, vignette code should not try to download from the internet, ect. Furthermore, many of the *Bioconductor* package specifics are distinct form the *CRAN* guidelines, which made finding suitable resources of knowledge even more limited. A straight forward guide for package creation would have been appreciated by me (a novice developer), and so I created subsection 2.3.1, as a potential resource for other novice developers. I believe this subsection of the thesis to be a particularly important contribution from this PhD thesis, and I plan on making this subsection available as a blog or short article in the near future. I will also contact the *Bioconductor* core team to flag this as a potential resource for them, because I do know they are interested in contributions from the developers community.

6.2 Ch3 and Ch4 - Chondrogenesis data analysis and creation of a Multi-miRNA chondrogenesis model

Results from Ch3 and Ch4 lead to the construction of a chondrogenesis based multi-miRNA kinetic model. This model captured the inherit complexity of a segment of TGFB induced chondrogenesis and may help researchers in further experimental design. This model contains multiple miRNAs (*miR-199a-5p, miR-199b-5p* and *miR-140-5p*) and was centred around validatory data for a novel miRNA-mRNA pair (*miR-199b-5p-CAV1*). The model presented also allowed for predictions to be made on how the system reacts during *miR-199a-5p* inhibition and predicts changes in *miR-140-5p* and GAG levels during *miR-199a-5p* or *miR-199b-5p* inhibition. Overall, this model has helped to establish *miR-199b-5p-CAV1* mRNA regulation as a possible regulatory interaction during chondrogenesis.

Ideally the validation data should be generated to test model predictions from a kinetic model calibrated on a comprehensive dataset. In practice this approach often proves difficult to implement, and qualitative strategies are employed. The more validatory data given to a model, the better the model can function to make predictions and capture the

inherit complexities of a signalling pathway. Our collaborators provided high quality data, and based on the amount received, several assumptions were made:

- miRNAs do not form complexes with target mRNAs. I.e. no miRISC-mRNA complexes will form. The rationale for this was the data the model was based on spans over days, so quicker reactions which may take seconds-hours would not be useful in this model.
- miRNAs are not degraded with their target mRNAs, and miRNA levels are not modulated by their mRNA targets at all.
- *miR-199a-5p* and *miR-199b-5p* double inhibition will lead to a very large affect on chondrogenesis, specifically a large increase in *CAV1* and a large decrease in chondrogenic biomarkers. I believe the model does not handle the double inhibition well. It is too extreme and unlikely.
- GAG levels were increasing over the time course, and seemed to plateau from around day 10. This may be inaccurate, but we only had GAG level information from day 7.
- during *miR-199b-5p* inhibition GAG levels decreased to 40% on day 7 based on results from Figure 4.3D.
- TGFB3 levels decreased at a steady rate during the 14 day time course.
- miRNA inhibition lasted for 5.5 days, after this time the drug effects wore off. This may be inaccurate, but I justified this because during *miR-199b-5p* inhibition, by day 7 the chondrogenic biomarkers began aligning to their calibration levels (Figure 4.7).
- Other miRNA target genes regulated chondrogenesis earlier and at a greater rate than *CAV1* mRNA. This was because *CAV1* mRNA levels peaked at day 3 after *hsa-miR-199b-5p* inhibition and *ACAN*, *COL2A1* and *SOX9* mRNA levels had their greatest nadir at day 1 after *hsa-miR-199b-5p* inhibition.
- The OtherTargets ”blackbox” represented all other mRNA targets other than *CAV1*.

Other *miR-199a/b-5p* targets

Within the model presented in Ch4, the OtherTargets species represented other *miR-199a/b-5p* targets (except *CAV1* mRNA). The nature of miRNA-mRNA interactions are complex and there could be several other *miR-199a/b-5p* targets which may regulate chondrogenesis. It is clear from the validatory data, alternative targets regulated chondrogenesis during or prior to day 1 of chondrogenesis (Figure 4.3A-C, Figure 4.4A). The positive results seen on Figure 4.3 prompted my collaborators to generate *miR-199a-5p* and *miR-199b-5p* inhibition RNAseq data measured at D0 and D1 after *miR-199a-5p* or *miR-199b-5p* inhibition to identify other target genes.

RNAseq - *miR-199a/b-5p* inhibition

Analysis of the RNAseq data found *CAV1* along with several other potential *miR-199a/b-5p* targets. The three targets with the lowest adjusted P values which were positively modulated during *miR-199a/b-5p* inhibition, in order were: *FZD6*, *ITGA3* and *CAV1*. Further qPCR analysis did show *FZD6* and *ITGA3* to be upregulated during *miR-199a-5p* and *miR-199b-5p* inhibition. Upon a literature search I found *FZD6* and *ITGA3* to have been anti-chondrogenic in function, but their roles within signalling pathways during chondrogenic regulation were not as well explored as the CAV1-RHoA/ROCK1 system [297, 374, 375]. *SkeletaVis* was used to check if *FZD6*, *ITGA3* and *CAV1* were all consistently negatively regulated in other chondrogenesis studies, and results seemed consistent with the RNAseq analysis [288]. It is likely that FZD6 and ITGA3 are involved in Wnt signalling and/or fibronectin generation. There is also some evidence that indicates FZD6 and CAV1 have some indirect interplay via RHoA activity during fibrillogenesis (generation of fibronectin within cells) [297, 376, 377]. I have created an updated model which incorporated *FZD6* and *ITGA3*.

This work was not shown in this thesis for three reasons:

1. Most importantly, I did not wish for this thesis to become encumbered with experimental work from my collaborators. This could lead to my own methods receiving less attention, and could also lead to this thesis not accurately reflecting my own

contributions to these projects.

2. The core theme of the thesis was the demonstration and exploration of multiple computational methodologies for the analysis of longitudinal miRNA-mRNA expression datasets. The RNAseq data generated by my collaborators had two time points, so would not count as a longitudinal miRNA-mRNA expression dataset. So I believed allocating a new chapter to discuss the RNAseq results would detract from the core message of the thesis.

3. Generation of GRNs and kinetic models from longitudinal miRNA-mRNA expression datasets was covered by Ch4, so adding more GRNs and another kinetic model may not have added any innovative content.

Further work

We intend to publish the updated model, along with the RNAseq data, RNAseq analysis and qRT-PCR validation data. I also intend to upload the updated model onto biomodels. It is an aim of reproducible modelling to have models at the center of experimental design, and the model generated during this PhD may help as a resource for experimental design by other chondrogenesis/ OA research groups in the future.

6.3 Ch5 - Machine Learning

Ch5 presented a ML project which I designed. Here a large longitudinal miRNA-mRNA expression dataset is explored using DE, *TimiRGeN* and ML. The JHD dataset analysed was the largest that I found, and consisted of 336 individual sequencing experiments. I co-supervised an UG student Bethany Harley. She contrasted JHD samples by gender and her work was the basis for Figure 5.1. Age was found to be the most important variable in the dataset, so this was further explored using ML. This project was an early predisposition approach. SDEGs found in both 6M and 10M data were treated as the training data, and the same SDEGs from the 2M data were treated as the testing data. Thus, we tried to identify a set of miRNAs and mRNAs which indicated predisposition to JHD, which currently has no biomarkers other than an extreme CAG-codon expansion in the *HTT* gene

which could be identified through screening.

I created a Logistic Regression model which identified WT samples with 100% accuracy, but the HD samples were only accurately identified for 74% of the samples (Figure 5.4A). Also, the ROC/AUC curve (Figure 5.4B) scored 87.1%. Further work must be done, however, this is a positive start. I have suspicions about the 6M data and it may be interesting to repeat this work without the 6M data, though this would significantly decrease our trainable samples.

Further work

The aim of the early predisposition project was to identify genes which can be used to train a model that can predict HD or WT samples. Over a quarter of the HD samples were being misplaced, so perhaps other methods of feature identification should be used e.g. feature selection techniques.

The miRNAs found in the early detection and age separated ML models could be ratified by examining overexpressed miRNAs in HD patient fluid samples (blood plasma or CSF.) I plan to download and analyse data from several public datasets to ratify the miRNAs which were presented as biomarkers for JHD.

CHAPTER 7

CONCLUSIONS AND BIBLIOGRAPHY

Longitudinal miRNA-mRNA expression datasets is a popular resource for biological investigation. A great range of insight can be gained from such complex datasets, however only if computational biologists have techniques and tools sophisticated enough to analyse longitudinal miRNA-mRNA datasets. In this PhD I have shown the use of three distinct techniques: the *TimiRGeN R* package which uses a big data/ bioinformatics approach to integrate, analyse and generate networks from longitudinal miRNA-mRNA datasets, a mutli-miRNA chondrogenesis model which uses principles from systems biology to establish how *miR-199b-5p* regulates chondrogenesis, and a ML approach which uses a large dataset to create an early predisposition model for JHD.

I developed the *TimiRGeN R/Bioconductor* package. This is the core achievement of the PhD because, not only was it accepted onto *Bioconductor*, but also it became the foundation for a first author original paper in *Bioinformatics* [117]. It is my hope that this tool helps other researchers find direction when analysing longitudinal miRNA-mRNA expression datasets. Also, I was invited to talk at Bioc2021, so this tool has been recognised by the *Bioconductor* community.

I constructed GRNs and a multi-miRNA chondrogenesis kinetic model from output of *TimiRGeN*. This model is centered around the novel interaction between *miR-199b-5p*

and *CAV1* mRNA. In collaboration with experimentalists, this model was validated and made several predictions, including predicting how *miR-199a-5p* may be regulating chondrogenesis. I will work on publishing this work to showcase how bioinformatics, kinetic modelling and wet-lab was used in combination to identify *miR-199a/b-5p* as novel regulators of chondrogenesis.

I designed an early predisposition ML project using a large JHD dataset. With help from project students, we have made a positive start by creating a logistic regression model with 100% accuracy at detecting WT samples, but more work is needed, as the HD detection rate is poor. I aim to publish this work to show how ML could be used to analyse large longitudinal miRNA-mRNA expression datasets.

Overall, I have utilised multiple computational methods to analyse several longitudinal miRNA-mRNA datasets during the last four years. I believe this PhD has contributed to the miRNA/ non-coding research community, especially in terms of the *TimiRGeN R* package, which has been downloaded by over 300 individual IP addresses since acceptance into *Bioconductor*. Using the *TimiRGeN R* package I identified *hsa-miR-199b-5p* to be an interesting novel miRNA in chondrogenesis. In collaboration with experimentalists from the Young group at Newcastle University, we have used techniques from systems biology to create a validated multi-miRNA chondrogenesis kinetic model. The GRNs and modelling helped in experimental design and results revealed *hsa-miR-199b-5p* to be an important pro-chondrogenic regulator. This work proves the usefulness of the *TimiRGeN R* package in hypothesis generation and also provides an example of how systems biology can help to make sense of the complex nature of miRNA-mRNA interactions. Finally, I have developed a ML model to determine if samples are WT or HD. This model is trained on 6M and 10M data and tested on 2M data, making an early predisposition to JHD detection model.

BIBLIOGRAPHY

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

[2] Y. Zeng, R. Yi, and B. R. Cullen, "Micrornas and small interfering rnas can inhibit mrna expression by similar mechanisms," *Proceedings of the National Academy of Sciences*, vol. 100, no. 17, pp. 9779–9784, 2003.

[3] C. Sevignani, G. A. Calin, L. D. Siracusa, and C. M. Croce, "Mammalian micrornas: a small world for fine-tuning gene expression," *Mammalian genome*, vol. 17, no. 3, pp. 189–202, 2006.

[4] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microrna targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[5] J. G. Doench and P. A. Sharp, "Specificity of microrna target selection in translational repression," *Genes & development*, vol. 18, no. 5, pp. 504–511, 2004.

[6] E. C. Lai, "Micro rnas are complementary to 3 utr sequence motifs that mediate negative post-transcriptional regulation," *Nature genetics*, vol. 30, no. 4, pp. 363–364, 2002.

[7] Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim, "Microrna genes are transcribed by rna polymerase ii," *The EMBO journal*, vol. 23, no. 20, pp. 4051–4060, 2004.

[8] G. M. Borchert, W. Lanier, and B. L. Davidson, “Rna polymerase iii transcribes human micrornas,” *Nature structural & molecular biology*, vol. 13, no. 12, pp. 1097–1101, 2006.

[9] J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, and V. N. Kim, “The drosha-dgcr8 complex in primary microrna processing,” *Genes & development*, vol. 18, no. 24, pp. 3016–3027, 2004.

[10] T. Fukuda, K. Yamagata, S. Fujiyama, T. Matsumoto, I. Koshida, K. Yoshimura, M. Mihara, M. Naitou, H. Endoh, T. Nakamura, *et al.*, “Dead-box rna helicase subunits of the drosha complex are required for processing of rrna and a subset of microrna,” *Nature cell biology*, vol. 9, no. 5, pp. 604–611, 2007.

[11] R. Yi, Y. Qin, I. G. Macara, and B. R. Cullen, “Exportin-5 mediates the nuclear export of pre-microrna and short hairpin rnas,” *Genes & development*, vol. 17, no. 24, pp. 3011–3016, 2003.

[12] E. Lund and J. Dahlberg, “Substrate selectivity of exportin 5 and dicer in the biogenesis of microrna,” in *Cold Spring Harbor symposia on quantitative biology*, vol. 71, pp. 59–66, Cold Spring Harbor Laboratory Press, 2006.

[13] V. N. Kim, “Microrna precursors in motion: exportin-5 mediates their nuclear export,” *Trends in cell biology*, vol. 14, no. 4, pp. 156–159, 2004.

[14] A. Tsutsumi, T. Kawamata, N. Izumi, H. Seitz, and Y. Tomari, “Recognition of the pre-mirna structure by drosophila dicer-1,” *Nature structural & molecular biology*, vol. 18, no. 10, p. 1153, 2011.

[15] T. P. Chendrimada, R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar, “Trbp recruits the dicer complex to ago2 for microrna processing and gene silencing,” *Nature*, vol. 436, no. 7051, pp. 740–744, 2005.

[16] L. Guo and Z. Lu, “The fate of mirna* strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule?,” *PloS one*, vol. 5, no. 6, 2010.

[17] J. Liu, M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J.-J. Song, S. M. Hammond, L. Joshua-Tor, and G. J. Hannon, "Argonaute2 is the catalytic engine of mammalian rnai," *Science*, vol. 305, no. 5689, pp. 1437–1441, 2004.

[18] J. S. Shapiro, R. A. Langlois, A. M. Pham, *et al.*, "Evidence for a cytoplasmic microprocessor of pri-mirnas," *Rna*, vol. 18, no. 7, pp. 1338–1346, 2012.

[19] E. Berezikov, W.-J. Chung, J. Willis, E. Cuppen, and E. C. Lai, "Mammalian mirtron genes," *Molecular cell*, vol. 28, no. 2, pp. 328–336, 2007.

[20] A. S. Flynt, J. C. Greimann, W.-J. Chung, C. D. Lima, and E. C. Lai, "Microrna biogenesis via splicing and exosome-mediated trimming in drosophila," *Molecular cell*, vol. 38, no. 6, pp. 900–907, 2010.

[21] C. Ender, A. Krek, M. R. Friedländer, M. Beitzinger, L. Weinmann, W. Chen, S. Pfeffer, N. Rajewsky, and G. Meister, "A human snorna with microrna-like functions," *Molecular cell*, vol. 32, no. 4, pp. 519–528, 2008.

[22] D. Haussecker, Y. Huang, A. Lau, P. Parameswaran, A. Z. Fire, and M. A. Kay, "Human trna-derived small rnas in the global regulation of rna silencing," *Rna*, vol. 16, no. 4, pp. 673–695, 2010.

[23] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl, "Human argonaute2 mediates rna cleavage targeted by mirnas and sirnas," *Molecular cell*, vol. 15, no. 2, pp. 185–197, 2004.

[24] S. Niaz, "The ago proteins: an overview," *Biological chemistry*, vol. 399, no. 6, pp. 525–547, 2018.

[25] L. Li, D. Zhu, L. Huang, J. Zhang, Z. Bian, X. Chen, Y. Liu, C.-Y. Zhang, and K. Zen, "Argonaute 2 complexes selectively protect the circulating microrna in cell-secreted microvesicles," *PloS one*, vol. 7, no. 10, 2012.

[26] E. R. Kingston and D. P. Bartel, "Global analyses of the dynamics of mammalian microrna metabolism," *Genome research*, vol. 29, no. 11, pp. 1777–1790, 2019.

[27] A. Eulalio, F. Tritschler, and E. Izaurralde, "The gw182 protein family in animal cells: new insights into domains required for mirna-mediated gene silencing," *Rna*, vol. 15, no. 8, pp. 1433–1442, 2009.

[28] J. Sheu-Gruttadauria and I. J. MacRae, "Phase transitions in the assembly and function of human mirisc," *Cell*, vol. 173, no. 4, pp. 946–957, 2018.

[29] J. T. Zipprich, S. Bhattacharyya, H. Mathys, and W. Filipowicz, "Importance of the c-terminal domain of the human gw182 protein tnrc6c for translational repression," *Rna*, vol. 15, no. 5, pp. 781–793, 2009.

[30] D. R. Gallie, "The cap and poly (a) tail function synergistically to regulate mrna translational efficiency.," *Genes & development*, vol. 5, no. 11, pp. 2108–2116, 1991.

[31] M. Derry, A. Yanagiya, Y. Martineau, and N. Sonenberg, "Regulation of poly (a)-binding protein through pabp-interacting proteins," in *Cold Spring Harbor symposia on quantitative biology*, vol. 71, pp. 537–543, Cold Spring Harbor Laboratory Press, 2006.

[32] L. Zekri, E. Huntzinger, S. Heimstädt, and E. Izaurralde, "The silencing domain of gw182 interacts with pabpc1 to promote translational repression and degradation of microrna targets and is required for target release," *Molecular and cellular biology*, vol. 29, no. 23, pp. 6220–6231, 2009.

[33] T. Fatscher, V. Boehm, B. Weiche, and N. H. Gehring, "The interaction of cytoplasmic poly (a)-binding protein with eukaryotic initiation factor 4g suppresses nonsense-mediated mrna decay," *Rna*, vol. 20, no. 10, pp. 1579–1592, 2014.

[34] D. T. Humphreys, B. J. Westman, D. I. Martin, and T. Preiss, "Micrornas control translation initiation by inhibiting eukaryotic initiation factor 4e/cap and poly (a) tail function," *Proceedings of the National Academy of Sciences*, vol. 102, no. 47, pp. 16961–16966, 2005.

[35] E. Huntzinger, J. E. Braun, S. Heimstädt, L. Zekri, and E. Izaurralde, "Two pabpc1-binding sites in gw182 proteins promote mirna-mediated gene silencing," *The EMBO journal*, vol. 29, no. 24, pp. 4146–4160, 2010.

[36] J. Lowell, D. Rudner, and A. Sachs, "3'-utr-dependent deadenylation by the yeast poly (a) nuclease.," *Genes & Development*, vol. 6, no. 11, pp. 2088–2099, 1992.

[37] R. Boeck, S. Tarun, M. Rieger, J. A. Deardorff, S. Müller-Auer, and A. B. Sachs, "The yeast pan2 protein is required for poly (a)-binding protein-stimulated poly (a)-nuclease activity," *Journal of Biological Chemistry*, vol. 271, no. 1, pp. 432–438, 1996.

[38] C. E. Brown, S. Tarun, R. Boeck, and A. B. Sachs, "Pan3 encodes a subunit of the pab1p-dependent poly (a) nuclease in saccharomyces cerevisiae.," *Molecular and cellular biology*, vol. 16, no. 10, pp. 5744–5753, 1996.

[39] A. Yamashita, T.-C. Chang, Y. Yamashita, W. Zhu, Z. Zhong, C.-Y. A. Chen, and A.-B. Shyu, "Concerted action of poly (a) nucleases and decapping enzyme in mammalian mrna turnover," *Nature structural & molecular biology*, vol. 12, no. 12, pp. 1054–1063, 2005.

[40] T. K. Albert, M. Lemaire, N. L. van Berkum, R. Gentz, M. A. Collart, and H. T. M. Timmers, "Isolation and characterization of human orthologs of yeast ccr4–not complex subunits," *Nucleic acids research*, vol. 28, no. 3, pp. 809–817, 2000.

[41] M. Tucker, M. A. Valencia-Sanchez, R. R. Staples, J. Chen, C. L. Denis, and R. Parker, "The transcription factor associated ccr4 and caf1 proteins are components of the major cytoplasmic mrna deadenylase in saccharomyces cerevisiae," *Cell*, vol. 104, no. 3, pp. 377–386, 2001.

[42] C. Temme, S. Zaessinger, S. Meyer, M. Simonelig, and E. Wahle, "A complex containing the ccr4 and caf1 proteins is involved in mrna deadenylation in drosophila," *The EMBO journal*, vol. 23, no. 14, pp. 2862–2871, 2004.

[43] I. Behm-Ansmant, J. Rehwinkel, T. Doerks, A. Stark, P. Bork, and E. Izaurralde, "mrna degradation by mirnas and gw182 requires both ccr4: Not deadenylase and dcp1: Dcp2 decapping complexes," *Genes & development*, vol. 20, no. 14, pp. 1885–1898, 2006.

[44] J. E. Braun, E. Huntzinger, M. Fauser, and E. Izaurralde, "Gw182 proteins directly recruit cytoplasmic deadenylase complexes to mirna targets," *Molecular cell*, vol. 44, no. 1, pp. 120–133, 2011.

[45] S. Jonas and E. Izaurralde, "The role of disordered protein regions in the assembly of decapping complexes and rnp granules," *Genes & development*, vol. 27, no. 24, pp. 2628–2641, 2013.

[46] J. Rehwinkel, I. Behm-Ansmant, D. Gatfield, and E. Izaurralde, "A crucial role for gw182 and the dcp1: Dcp2 decapping complex in mirna-mediated gene silencing," *Rna*, vol. 11, no. 11, pp. 1640–1647, 2005.

[47] L. Wu, J. Fan, and J. G. Belasco, "Micrornas direct rapid deadenylation of mrna," *Proceedings of the National Academy of Sciences*, vol. 103, no. 11, pp. 4034–4039, 2006.

[48] S. Jonas and E. Izaurralde, "Towards a molecular understanding of microrna-mediated gene silencing," *Nature reviews genetics*, vol. 16, no. 7, pp. 421–433, 2015.

[49] F. Frank, N. Sonenberg, and B. Nagar, "Structural basis for 5-nucleotide base-specific recognition of guide rna by human ago2," *Nature*, vol. 465, no. 7299, pp. 818–822, 2010.

[50] T. Kawamata, M. Yoda, and Y. Tomari, "Multilayer checkpoints for microrna authenticity during risc assembly," *EMBO reports*, vol. 12, no. 9, pp. 944–949, 2011.

[51] D. S. Schwarz, G. Hutvágner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore, "Asymmetry in the assembly of the rnai enzyme complex," *Cell*, vol. 115, no. 2, pp. 199–208, 2003.

[52] J. C. Medley, G. Panzade, and A. Y. Zinovyeva, "microrna strand selection: Unwinding the rules," *Wiley Interdisciplinary Reviews: RNA*, vol. 12, no. 3, p. e1627, 2021.

[53] S. L. Ameres, M. D. Horwich, J.-H. Hung, J. Xu, M. Ghildiyal, Z. Weng, and P. D. Zamore, "Target rna–directed trimming and tailing of small silencing rnas," *Science*, vol. 328, no. 5985, pp. 1534–1539, 2010.

[54] A. Bitetti, A. C. Mallory, E. Golini, C. Carrieri, H. C. Gutierrez, E. Perlas, Y. A. Perez-Rico, G. P. Tocchini-Valentini, A. J. Enright, W. H. Norton, *et al.*, "Microrna degradation by a conserved target rna regulates animal behavior," *Nature structural & molecular biology*, vol. 25, no. 3, pp. 244–251, 2018.

[55] J. Sheu-Gruttadauria, P. Pawlica, S. M. Klum, S. Wang, T. A. Yario, N. T. S. Oakdale, J. A. Steitz, and I. J. MacRae, "Structural basis for target-directed microrna degradation," *Molecular cell*, vol. 75, no. 6, pp. 1243–1255, 2019.

[56] J. Zhang, S. Li, L. Li, M. Li, C. Guo, J. Yao, and S. Mi, "Exosome and exosomal microrna: trafficking, sorting, and function," *Genomics, proteomics & bioinformatics*, vol. 13, no. 1, pp. 17–24, 2015.

[57] J. Guduric-Fuchs, A. OConnor, B. Camp, C. L. O'Neill, R. J. Medina, and D. A. Simpson, "Selective extracellular vesicle-mediated export of an overlapping set of micrornas from multiple cell types," *BMC genomics*, vol. 13, no. 1, pp. 1–14, 2012.

[58] S. Kumar, M. Vijayan, J. Bhatti, and P. Reddy, "Micrornas as peripheral biomarkers in aging and age-related diseases," *Progress in molecular biology and translational science*, vol. 146, pp. 47–94, 2017.

[59] S. M. Peterson, J. A. Thompson, M. L. Ufkin, P. Sathyanarayana, L. Liaw, and C. B. Congdon, "Common features of microrna target prediction tools," *Frontiers in genetics*, vol. 5, p. 23, 2014.

[60] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, "Most mammalian mrnas are conserved targets of micrornas," *Genome research*, vol. 19, no. 1, pp. 92–105, 2009.

[61] D. Yue, H. Liu, and Y. Huang, "Survey of computational algorithms for microrna target prediction," *Current genomics*, vol. 10, no. 7, pp. 478–492, 2009.

[62] D. Long, R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding, "Potent effect of target structure on microrna function," *Nature structural & molecular biology*, vol. 14, no. 4, pp. 287–294, 2007.

[63] D. M. Garcia, D. Baek, C. Shin, G. W. Bell, A. Grimson, and D. P. Bartel, "Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other micrornas," *Nature structural & molecular biology*, vol. 18, no. 10, p. 1139, 2011.

[64] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel, "Microrna targeting specificity in mammals: determinants beyond seed pairing," *Molecular cell*, vol. 27, no. 1, pp. 91–105, 2007.

[65] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, "Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites," *Genome biology*, vol. 11, no. 8, p. R90, 2010.

[66] Z. Fang and N. Rajewsky, "The impact of mirna target sites in coding sequences and in 3 utrs," *PloS one*, vol. 6, no. 3, p. e18067, 2011.

[67] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microrna target recognition," *Nature genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[68] D. C. Ellwanger, F. A. Büttner, H.-W. Mewes, and V. Stümpflen, "The sufficient minimal set of mirna seed types," *Bioinformatics*, vol. 27, no. 10, pp. 1346–1350, 2011.

[69] G. Pio, M. Ceci, D. Malerba, and D. D'Elia, "Comirnet: a web-based system for the analysis of mirna-gene regulatory networks," *BMC bioinformatics*, vol. 16, no. 9, pp. 1–18, 2015.

[70] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "mirecords: an integrated resource for microrna–target interactions," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D105–D110, 2009.

[71] H. Naeem, R. Küffner, G. Csaba, and R. Zimmer, "mirsel: automated extraction of associations between microrna and genes from the biomedical literature," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–8, 2010.

[72] H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, Y. Tang, Y.-G. Chen, C.-N. Jin, Y. Yu, *et al.*, "mirtarbase 2020: updates to the experimentally validated microrna–target interaction database," *Nucleic acids research*, vol. 48, no. D1, pp. D148–D154, 2020.

[73] C. Sticht, C. De La Torre, A. Parveen, and N. Gretz, "mirwalk: An online resource for prediction of microrna binding sites," *PloS one*, vol. 13, no. 10, p. e0206239, 2018.

[74] Z.-W. Guo, C. Xie, J.-R. Yang, J.-H. Li, J.-H. Yang, and L. Zheng, "Mtibase: a database for decoding microrna target sites located within cds and 5 utr regions from clip-seq and expression profile datasets," *Database*, vol. 2015, 2015.

[75] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data," *Nucleic acids research*, vol. 42, no. D1, pp. D92–D97, 2014.

[76] D. Karagkouni, M. D. Paraskevopoulou, S. Chatzopoulos, I. S. Vlachos, S. Tastsoglou, I. Kanellos, D. Papadimitriou, I. Kavakiotis, S. Maniou, G. Skoufos, *et al.*, "Diana-tarbase v8: a decade-long collection of experimentally supported mirna–gene interactions," *Nucleic acids research*, vol. 46, no. D1, pp. D239–D245, 2018.

[77] N. K. Singh, "mirnas target databases: developmental methods and target identification techniques with functional annotations," *Cellular and Molecular Life Sciences*, vol. 74, no. 12, pp. 2239–2261, 2017.

[78] V. A. Gennarino, M. Sardiello, M. Mutarelli, G. Dharmalingam, V. Maselli, G. Lago, and S. Banfi, "Hoctar database: a unique resource for microrna target prediction," *Gene*, vol. 480, no. 1-2, pp. 51–58, 2011.

[79] Y. Chen and X. Wang, "mirdb: an online database for prediction of functional microrna targets," *Nucleic acids research*, vol. 48, no. D1, pp. D127–D131, 2020.

[80] J. Piriyapongsa, C. Bootchai, C. Ngamphiw, and S. Tongsima, "micropir: an integrated database of microrna target sites within human promoter sequences," *PloS one*, vol. 7, no. 3, p. e33888, 2012.

[81] Y. Ru, K. J. Kechris, B. Tabakoff, P. Hoffman, R. A. Radcliffe, R. Bowler, S. Mahaffey, S. Rossi, G. A. Calin, L. Bemis, *et al.*, "The multimir r package and database: integration of microrna–target interactions along with their disease and drug associations," *Nucleic acids research*, vol. 42, no. 17, pp. e133–e133, 2014.

[82] J. L. Rukov, R. Wilentzik, I. Jaffe, J. Vinther, and N. Shomron, "Pharmaco-mir: linking micrornas and drug effects," *Briefings in bioinformatics*, vol. 15, no. 4, pp. 648–659, 2014.

[83] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microrna target sites in mammalian mrnas," *elife*, vol. 4, p. e05005, 2015.

[84] A. Kozomara and S. Griffiths-Jones, "mirbase: annotating high confidence micrornas using deep sequencing data," *Nucleic acids research*, vol. 42, no. D1, pp. D68–D73, 2014.

[85] K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel, "The widespread impact of mammalian micrornas on mrna repression and evolution," *Science*, vol. 310, no. 5755, pp. 1817–1821, 2005.

[86] C. P. Bracken, J. M. Szubert, T. R. Mercer, M. E. Dinger, D. W. Thomson, J. S. Mattick, M. Z. Michael, and G. J. Goodall, "Global analysis of the mammalian rna degradome reveals widespread mirna-dependent and mirna-independent endonucleolytic cleavage," *Nucleic acids research*, vol. 39, no. 13, pp. 5658–5668, 2011.

[87] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, "Mapping the human mirna interactome by clash reveals frequent noncanonical binding," *Cell*, vol. 153, no. 3, pp. 654–665, 2013.

[88] S. M. Hartig, M. P. Hamilton, D. A. Bader, and S. E. McGuire, "The mirna interactome in metabolic homeostasis," *Trends in Endocrinology & Metabolism*, vol. 26, no. 12, pp. 733–745, 2015.

[89] J. B. Lian, G. S. Stein, A. J. Van Wijnen, J. L. Stein, M. Q. Hassan, T. Gaur, and Y. Zhang, "Microrna control of bone formation and homeostasis," *Nature Reviews Endocrinology*, vol. 8, no. 4, p. 212, 2012.

[90] K. Chandrasekaran, D. S. Karolina, S. Sepramaniam, A. Armugam, E. M. Wintour, J. F. Bertram, and K. Jeyaseelan, "Role of microrna s in kidney homeostasis and disease," *Kidney international*, vol. 81, no. 7, pp. 617–627, 2012.

[91] S. Miyaki, T. Sato, A. Inoue, S. Otsuki, Y. Ito, S. Yokoyama, Y. Kato, F. Takemoto, T. Nakasa, S. Yamashita, *et al.*, "Microrna-140 plays dual roles in both cartilage development and homeostasis," *Genes & development*, vol. 24, no. 11, pp. 1173–1185, 2010.

[92] G. A. Calin, C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, *et al.*, "Frequent deletions and down-regulation of micro-rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia," *Proceedings of the national academy of sciences*, vol. 99, no. 24, pp. 15524–15529, 2002.

[93] S. D. Linnstaedt, E. Gottwein, R. L. Skalsky, M. A. Luftig, and B. R. Cullen, "Virally induced cellular microrna mir-155 plays a key role in b-cell immortalization by epstein-barr virus," *Journal of virology*, vol. 84, no. 22, pp. 11670–11678, 2010.

[94] M. Trajkovski, J. Hausser, J. Soutschek, B. Bhat, A. Akin, M. Zavolan, M. H. Heim, and M. Stoffel, "Micrornas 103 and 107 regulate insulin sensitivity," *Nature*, vol. 474, no. 7353, pp. 649–653, 2011.

[95] J. Kim, K. Inoue, J. Ishii, W. B. Vanti, S. V. Voronov, E. Murchison, G. Hannon, and A. Abeliovich, "A microrna feedback circuit in midbrain dopamine neurons," *Science*, vol. 317, no. 5842, pp. 1220–1224, 2007.

[96] H. Valadi, K. Ekström, A. Bossios, M. Sjöstrand, J. J. Lee, and J. O. Lötvall, "Exosome-mediated transfer of mrnas and micrornas is a novel mechanism of genetic exchange between cells," *Nature cell biology*, vol. 9, no. 6, pp. 654–659, 2007.

[97] V. Swarup and M. Rajeswari, "Circulating (cell-free) nucleic acids–a promising, non-

invasive tool for early detection of several human diseases," *FEBS letters*, vol. 581, no. 5, pp. 795–799, 2007.

[98] C. H. Lawrie, S. Gal, H. M. Dunlop, B. Pushkaran, A. P. Liggins, K. Pulford, A. H. Banham, F. Pezzella, J. Boultwood, J. S. Wainscoat, *et al.*, "Detection of elevated levels of tumour-associated microrna in serum of patients with diffuse large b-cell lymphoma," *British journal of haematology*, vol. 141, no. 5, pp. 672–675, 2008.

[99] S. S. Chim, T. K. Shing, E. C. Hung, T.-y. Leung, T.-k. Lau, R. W. Chiu, and Y. Dennis Lo, "Detection and characterization of placental microrna in maternal plasma," *Clinical chemistry*, vol. 54, no. 3, pp. 482–490, 2008.

[100] K. Wang, S. Zhang, B. Marzolf, P. Troisch, A. Brightman, Z. Hu, L. E. Hood, and D. J. Galas, "Circulating microrna, potential biomarkers for drug-induced liver injury," *Proceedings of the National Academy of Sciences*, vol. 106, no. 11, pp. 4402–4407, 2009.

[101] R. Silakit, W. Loilome, P. Yongvanit, S. Thongchot, P. Sithithaworn, T. Boonmars, S. Koonmee, A. Titapun, N. Khuntikeo, N. Chamadol, *et al.*, "Urinary microrna-192 and microrna-21 as potential indicators for liver fluke-associated cholangiocarcinoma risk group," *Parasitology international*, vol. 66, no. 4, pp. 479–485, 2017.

[102] C. Sun, N. Li, Z. Yang, B. Zhou, Y. He, D. Weng, Y. Fang, P. Wu, P. Chen, X. Yang, *et al.*, "mir-9 regulation of brca1 and ovarian cancer sensitivity to cisplatin and parp inhibition," *Journal of the National Cancer Institute*, vol. 105, no. 22, pp. 1750–1758, 2013.

[103] P. Moskwa, F. M. Buffa, Y. Pan, R. Panchakshari, P. Gottipati, R. J. Muschel, J. Beech, R. Kulshrestha, K. Abdelmohsen, D. M. Weinstock, *et al.*, "mir-182-mediated downregulation of brca1 impacts dna repair and sensitivity to parp inhibitors," *Molecular cell*, vol. 41, no. 2, pp. 210–220, 2011.

[104] T. A. Karlsen, G. A. de Souza, B. Ødegaard, L. Engebretsen, and J. E. Brinchmann, "microrna-140 inhibits inflammation and stimulates chondrogenesis in a model of interleukin 1β-induced osteoarthritis," *Molecular Therapy-Nucleic Acids*, vol. 5, p. e373, 2016.

[105] E. Ntoumou, M. Tzetis, M. Braoudaki, G. Lambrou, M. Poulou, K. Malizos, N. Stefanou, L. Anastasopoulou, and A. Tsezou, "Serum microrna array analysis identifies mir-140-3p, mir-33b-3p and mir-671-3p as potential osteoarthritis biomarkers involved in metabolic processes," *Clinical epigenetics*, vol. 9, no. 1, p. 127, 2017.

[106] E. Araldi and E. Schipani, "Microrna-140 and the silencing of osteoarthritis," *Genes & development*, vol. 24, no. 11, pp. 1075–1080, 2010.

[107] E. Clough and T. Barrett, "The gene expression omnibus database," in *Statistical genomics*, pp. 93–110, Springer, 2016.

[108] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, *et al.*, "Arrayexpress update–from bulk to single-cell expression data," *Nucleic acids research*, vol. 47, no. D1, pp. D711–D715, 2019.

[109] F. L. Craciun, V. Bijol, A. K. Ajay, P. Rao, R. K. Kumar, J. Hutchinson, O. Hofmann, N. Joshi, J. P. Luyendyk, U. Kusebauch, *et al.*, "Rna sequencing identifies novel translational biomarkers of kidney fibrosis," *Journal of the American Society of Nephrology*, vol. 27, no. 6, pp. 1702–1713, 2016.

[110] K. L. Pellegrini, C. V. Gerlach, F. L. Craciun, K. Ramachandran, V. Bijol, H. T. Kissick, and V. S. Vaidya, "Application of small rna sequencing to identify micrornas in acute kidney injury and fibrosis," *Toxicology and applied pharmacology*, vol. 312, pp. 42–52, 2016.

[111] J. Baran-Gale, J. E. Purvis, and P. Sethupathy, "An integrative transcriptomics approach identifies mir-503 as a candidate master regulator of the estrogen response in mcf-7 breast cancer cells," *Rna*, vol. 22, no. 10, pp. 1592–1603, 2016.

[112] C. Camps, H. K. Saini, D. R. Mole, H. Choudhry, M. Reczko, J. A. Guerra-Assunção, Y.-M. Tian, F. M. Buffa, A. L. Harris, A. G. Hatzigeorgiou, *et al.*, "Integrated analysis of microrna and mrna expression and association with hif binding reveals the complexity of microrna expression regulation under hypoxia," *Molecular cancer*, vol. 13, no. 1, p. 28, 2014.

[113] M. J. Barter, M. Tselepi, R. Gómez, S. Woods, W. Hui, G. R. Smith, D. P. Shanley, I. M. Clark, and D. A. Young, "Genome-wide microrna and gene analysis of mesenchymal stem cell chondrogenesis identifies an essential role and multiple targets for mir-140-5p," *Stem Cells*, vol. 33, no. 11, pp. 3266–3280, 2015.

[114] P. Langfelder, J. P. Cantle, D. Chatzopoulou, N. Wang, F. Gao, I. Al-Ramahi, X.-H. Lu, E. M. Ramos, K. El-Zein, Y. Zhao, *et al.*, "Integrated genomics and proteomics define huntingtin cag length–dependent networks in mice," *Nature neuroscience*, vol. 19, no. 4, p. 623, 2016.

[115] P. Langfelder, F. Gao, N. Wang, D. Howland, S. Kwak, T. F. Vogt, J. S. Aaronson, J. Rosinski, G. Coppola, S. Horvath, *et al.*, "Microrna signatures of endogenous huntingtin cag repeat expansion in mice," *PloS one*, vol. 13, no. 1, p. e0190550, 2018.

[116] D. Spies, P. F. Renz, T. A. Beyer, and C. Ciaudo, "Comparative analysis of differential gene expression tools for rna sequencing time course data," *Briefings in bioinformatics*, vol. 20, no. 1, pp. 288–298, 2019.

[117] K. Patel, S. Chandrasegaran, I. Clark, C. Proctor, D. Young, and D. Shanley, "Timirgen: R/bioconductor package for time series microrna–mrna integration and analysis," *Bioinformatics*, 2021.

[118] B. P. Ingalls, *Mathematical modeling in systems biology: an introduction*. MIT press, 2013.

[119] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer, "Copasia complex pathway simulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, 2006.

[120] M. P. van Iersel, T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin, and C. Evelo, "Presenting and exploring biological pathways with pathvisio," *BMC bioinformatics*, vol. 9, no. 1, p. 399, 2008.

[121] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8:

new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2010.

[122] M. Zhang, S. Z. Yuan, H. Sun, L. Sun, D. Zhou, and J. Yan, "mir-199b-5p promoted chondrogenic differentiation of c3h10t1/2 cells by regulating jag1," *Journal of Tissue Engineering and Regenerative Medicine*, vol. 14, no. 11, pp. 1618–1629, 2020.

[123] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The kegg resource for deciphering the genome," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D277–D280, 2004.

[124] D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. Digles, *et al.*, "Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research," *Nucleic acids research*, vol. 46, no. D1, pp. D661–D667, 2018.

[125] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[126] D. Diaz and S. Draghici, "mirintegrator: Integrating mirnas into signaling pathways," *R package*, 2015.

[127] T.-T. Wang, C.-Y. Lee, L.-C. Lai, M.-H. Tsai, T.-P. Lu, and E. Y. Chuang, "anamir: integrated analysis of microrna and gene expression profiling," *BMC bioinformatics*, vol. 20, no. 1, p. 239, 2019.

[128] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "Microrna targets in drosophila," *Genome biology*, vol. 5, no. 1, p. R1, 2003.

[129] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. Da Piedade, K. C. Gunsalus, M. Stoffel, *et al.*, "Combinatorial microrna target predictions," *Nature genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[130] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou, "A combined computational-experimental approach predicts

human microrna targets," *Genes & development*, vol. 18, no. 10, pp. 1165–1178, 2004.

[131] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, "Fast and effective prediction of microrna/target duplexes," *Rna*, vol. 10, no. 10, pp. 1507–1517, 2004.

[132] P. Maziere and A. J. Enright, "Prediction of microrna targets," *Drug discovery today*, vol. 12, no. 11-12, pp. 452–458, 2007.

[133] Y. Zhang and F. J. Verbeek, "Comparison and integration of target prediction algorithms for microrna studies," *Journal of integrative bioinformatics*, vol. 7, no. 3, pp. 169–181, 2010.

[134] C. Cava, A. Colaprico, G. Bertoli, A. Graudenzi, T. C. Silva, C. Olsen, H. Noushmehr, G. Bontempi, G. Mauri, and I. Castiglioni, "Spidermir: an r/bioconductor package for integrative analysis with mirna data," *International journal of molecular sciences*, vol. 18, no. 2, p. 274, 2017.

[135] M. Vila-Casadesús, M. Gironella, and J. J. Lozano, "Mircomb: an r package to analyse mirna-mrna interactions. examples across five digestive cancers," *PloS one*, vol. 11, no. 3, p. e0151127, 2016.

[136] C. J. Creighton, A. K. Nagaraja, S. M. Hanash, M. M. Matzuk, and P. H. Gunaratne, "A bioinformatics tool for linking gene expression profiling results with public databases of microrna target predictions," *Rna*, vol. 14, no. 11, pp. 2290–2296, 2008.

[137] Y. Fan and J. Xia, "mirnetfunctional analysis and visual exploration of mirna–target interactions in a network context," in *Computational cell biology*, pp. 215–233, Springer, 2018.

[138] S. L'Yi, D. Jung, M. Oh, B. Kim, R. J. Freishtat, M. Giri, E. Hoffman, and J. Seo, "mirtarvis+: Web-based interactive visual analytics tool for microrna target predictions," *Methods*, vol. 124, pp. 78–88, 2017.

[139] C. Wu, E. E. Bardes, A. G. Jegga, and B. J. Aronow, “Toppmir: ranking microrna and their mrna targets based on biological functions and context,” *Nucleic acids research*, vol. 42, no. W1, pp. W107–W113, 2014.

[140] A. Bisognin, G. Sales, A. Coppe, S. Bortoluzzi, and C. Romualdi, “Magia2: from mirna and genes expression data integrative analysis to microrna–transcription factor mixed regulatory circuits (2012 update),” *Nucleic acids research*, vol. 40, no. W1, pp. W13–W21, 2012.

[141] M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph, “Drem 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data,” *BMC systems biology*, vol. 6, no. 1, p. 104, 2012.

[142] E. Andrés-León, R. Núñez-Torres, and A. M. Rojas, “miarma-seq: a comprehensive tool for mirna, mrna and circrna analysis,” *Scientific reports*, vol. 6, p. 25749, 2016.

[143] M. J. Nueda, S. Tarazona, and A. Conesa, “Next masigpro: updating masigpro bioconductor package for rna-seq time series,” *Bioinformatics*, vol. 30, no. 18, pp. 2598–2602, 2014.

[144] A. Michna, H. Braselmann, M. Selmansberger, A. Dietz, J. Hess, M. Gomolka, S. Hornhardt, N. Blüthgen, H. Zitzelsberger, and K. Unger, “Natural cubic spline regression modeling followed by dynamic network reconstruction for the identification of radiation-sensitivity gene association networks from time-course transcriptome data,” *PloS one*, vol. 11, no. 8, p. e0160791, 2016.

[145] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for rna-seq data with deseq2,” *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[146] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for rna-sequencing and microarray studies,” *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[147] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edger: a bioconductor package

for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[148] Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "Continuous representations of time-series gene expression data," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 341–356, 2003.

[149] G. O. Consortium, "The gene ontology (go) project in 2006," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D322–D326, 2006.

[150] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 46, no. D1, pp. D649–D655, 2018.

[151] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, "Omnipath: guidelines and gateway for literature-curated signaling pathway resources," *Nature methods*, vol. 13, no. 12, pp. 966–967, 2016.

[152] L. Kumar and M. E. Futschik, "Mfuzz: a software package for soft clustering of microarray data," *Bioinformation*, vol. 2, no. 1, p. 5, 2007.

[153] G. Csardi, T. Nepusz, *et al.*, "The igraph software package for complex network research," *InterJournal, complex systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[154] J. A. Gustavsen, S. Pai, R. Isserlin, B. Demchak, and A. R. Pico, "Rcy3: Network biology using cytoscape from within r," *F1000Research*, vol. 8, 2019.

[155] J. Ding and Z. Bar-Joseph, "Analysis of time series regulatory networks," *Current Opinion in Systems Biology*, 2020.

[156] D. E. Jung, J. Wen, T. Oh, and S. Y. Song, "Differentially expressed microrna in pancreatic cancer stem cells," *Pancreas*, vol. 40, no. 8, pp. 1180–1187, 2011.

[157] U. Lakshmipathy, B. Love, L. A. Goff, R. Jörnsten, R. Graichen, R. P. Hart, and J. D. Chesnut, "Microrna expression pattern of undifferentiated and differentiated human embryonic stem cells," *Stem cells and development*, vol. 16, no. 6, pp. 1003–1016, 2007.

[158] V. Jayaswal, M. Lutherborrow, D. D. Ma, and Y. Hwa Yang, “Identification of micrornas with regulatory potential using a matched microrna-mrna time-course data,” *Nucleic acids research*, vol. 37, no. 8, pp. e60–e60, 2009.

[159] M. Szumilas, “Explaining odds ratios,” *Journal of the Canadian academy of child and adolescent psychiatry*, vol. 19, no. 3, p. 227, 2010.

[160] A. Chuasuwan and J. A. Kellum, “Acute kidney injury and its management,” in *Hemodialysis*, vol. 171, pp. 218–225, Karger Publishers, 2011.

[161] S. He, N. Liu, G. Bayliss, and S. Zhuang, “Egfr activity is required for renal tubular cell dedifferentiation and proliferation in a murine model of folic acid-induced acute kidney injury,” *American Journal of Physiology-Renal Physiology*, vol. 304, no. 4, pp. F356–F366, 2013.

[162] C. Chen, C. Lu, Y. Qian, H. Li, Y. Tan, L. Cai, and H. Weng, “Urinary mir-21 as a potential biomarker of hypertensive kidney injury and fibrosis,” *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[163] H.-W. Kao, C.-Y. Pan, C.-H. Lai, C.-W. Wu, W.-L. Fang, K.-H. Huang, and W.-C. Lin, “Urine mir-21-5p as a potential non-invasive biomarker for gastric cancer,” *Oncotarget*, vol. 8, no. 34, p. 56389, 2017.

[164] N. Ghorbanmehr, S. Gharbi, E. Korsching, M. Tavallaei, B. Einollahi, and S. J. Mowla, “mir-21-5p, mir-141-3p, and mir-205-5p levels in urinepromising biomarkers for the identification of prostate and bladder cancer,” *The Prostate*, vol. 79, no. 1, pp. 88–95, 2019.

[165] P. Tangtanatakul, S. Klinchanhom, P. Sodsai, T. Sutichet, C. Promjeen, Y. Avihingsanon, and N. Hirankarn, “Down-regulation of let-7a and mir-21 in urine exosomes from lupus nephritis patients during disease flare,” *Asian Pac. J. Allergy Immunol*, vol. 37, pp. 189–197, 2019.

[166] C. Liu, M. Wang, M. Chen, K. Zhang, L. Gu, Q. Li, Z. Yu, N. Li, and Q. Meng, “mir-18a induces myotubes atrophy by down-regulating igfi,” *The international journal of biochemistry & cell biology*, vol. 90, pp. 145–154, 2017.

[167] H. Liu, W. Chu, L. Gong, X. Gao, and W. Wang, "Microrna-26b is upregulated in a double transgenic mouse model of alzheimer's disease and promotes the expression of amyloid-β by targeting insulin-like growth factor 1," *Molecular Medicine Reports*, vol. 13, no. 3, pp. 2809–2814, 2016.

[168] Y.-K. Hu, X. Wang, L. Li, Y.-H. Du, H.-T. Ye, and C.-Y. Li, "Microrna-98 induces an alzheimers disease-like disturbance by targeting insulin-like growth factor 1," *Neuroscience bulletin*, vol. 29, no. 6, pp. 745–751, 2013.

[169] W. Sun, S. Hu, J. Hu, S. Yang, B. Hu, J. Qiu, X. Gan, H. Liu, L. Li, and J. Wang, "mir-365 inhibits duck myoblast proliferation by targeting igf-i via pi3k/akt pathway," *Bioscience Reports*, vol. 39, no. 11, 2019.

[170] G.-X. Liu, S. Ma, Y. Li, Y. Yu, Y.-X. Zhou, Y.-D. Lu, L. Jin, Z.-L. Wang, and J.-H. Yu, "Hsa-let-7c controls the committed differentiation of igf-1-treated mesenchymal stem cells derived from dental pulps by targeting igf-1r via the mapk pathways," *Experimental & molecular medicine*, vol. 50, no. 4, pp. 1–14, 2018.

[171] S. Roush and F. J. Slack, "The let-7 family of microrna," *Trends in cell biology*, vol. 18, no. 10, pp. 505–516, 2008.

[172] S. Gao, C. Cheng, H. Chen, M. Li, K. Liu, and G. Wang, "Igf1 3 utr functions as a cerna in promoting angiogenesis by sponging mir-29 family in osteosarcoma," *Journal of molecular histology*, vol. 47, no. 2, pp. 135–143, 2016.

[173] S. Subramanian, "Competing endogenous rnas (cernas): new entrants to the intricacies of gene regulation," *Frontiers in genetics*, vol. 5, p. 8, 2014.

[174] C. F. Hung, M. G. Rohani, S.-s. Lee, P. Chen, and L. M. Schnapp, "Role of igf-1 pathway in lung fibroblast activation," *Respiratory research*, vol. 14, no. 1, pp. 1–12, 2013.

[175] M. E. Futschik and L. Kumar, "Introduction to mfuzz package and its graphical user interface," 2013.

[176] F. Genovese, A. A. Manresa, D. J. Leeming, M. A. Karsdal, and P. Boor, "The extracellular matrix in the kidney: a source of novel non-invasive biomarkers of kidney fibrosis?," *Fibrogenesis & tissue repair*, vol. 7, no. 1, p. 4, 2014.

[177] X. Wen, Z. Peng, Y. Li, H. Wang, J. V. Bishop, L. R. Chedwick, K. Singbartl, and J. A. Kellum, "One dose of cyclosporine a is protective at initiation of folic acid-induced acute kidney injury in mice," *Nephrology Dialysis Transplantation*, vol. 27, no. 8, pp. 3100–3109, 2012.

[178] L. J. Stallons, R. M. Whitaker, and R. G. Schnellmann, "Suppressed mitochondrial biogenesis in folic acid-induced acute kidney injury and early fibrosis," *Toxicology letters*, vol. 224, no. 3, pp. 326–332, 2014.

[179] J. P. Wang and A. Hielscher, "Fibronectin: how its aberrant expression in tumors may improve therapeutic targeting," *Journal of Cancer*, vol. 8, no. 4, p. 674, 2017.

[180] A. J. Kriegel, Y. Liu, Y. Fang, X. Ding, and M. Liang, "The mir-29 family: genomics, cell biology, and relevance to renal and cardiovascular injury," *Physiological genomics*, vol. 44, no. 4, pp. 237–244, 2012.

[181] J. C. Broen, T. R. Radstake, and M. Rossato, "The role of genetics and epigenetics in the pathogenesis of systemic sclerosis," *Nature Reviews Rheumatology*, vol. 10, no. 11, p. 671, 2014.

[182] B. Su, W. Zhao, B. Shi, Z. Zhang, X. Yu, F. Xie, Z. Guo, X. Zhang, J. Liu, Q. Shen, *et al.*, "Let-7d suppresses growth, metastasis, and tumor macrophage infiltration in renal cell carcinoma by targeting col3a1 and ccl7," *Molecular cancer*, vol. 13, no. 1, p. 206, 2014.

[183] C.-M. Tang, M. Zhang, L. Huang, Z.-q. Hu, J.-N. Zhu, Z. Xiao, Z. Zhang, Q.-x. Lin, X.-L. Zheng, S.-L. Wu, *et al.*, "Circrna_000203 enhances the expression of fibrosis-associated genes by derepressing targets of mir-26b-5p, col1a2 and ctgf, in cardiac fibroblasts," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[184] M. Pavkovic, L. Pantano, C. V. Gerlach, S. Brutus, S. A. Boswell, R. A. Everley, J. V.

Shah, S. H. Sui, and V. S. Vaidya, “Multi omics analysis of fibrotic kidneys in two mouse models,” *Scientific data*, vol. 6, no. 1, pp. 1–9, 2019.

[185] E. Martínez-Klimova, O. E. Aparicio-Trejo, E. Tapia, and J. Pedraza-Chaverri, “Unilateral ureteral obstruction as a model to investigate fibrosis-attenuating treatments,” *Biomolecules*, vol. 9, no. 4, p. 141, 2019.

[186] R. C. Team *et al.*, “R: A language and environment for statistical computing,” 2013.

[187] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, *et al.*, “Orchestrating high-throughput genomic analysis with bioconductor,” *Nature methods*, vol. 12, no. 2, p. 115, 2015.

[188] R. Leinonen, H. Sugawara, M. Shumway, and I. N. S. D. Collaboration, “The sequence read archive,” *Nucleic acids research*, vol. 39, no. suppl_1, pp. D19–D21, 2010.

[189] S. Andrews *et al.*, “Fastqc: a quality control tool for high throughput sequence data,” 2010.

[190] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet. journal*, vol. 17, no. 1, pp. 10–12, 2011.

[191] F. Cunningham, P. Achuthan, W. Akanni, J. Allen, M. R. Amode, I. M. Armean, R. Bennett, J. Bhai, K. Billis, S. Boddu, *et al.*, “Ensembl 2019,” *Nucleic acids research*, vol. 47, no. D1, pp. D745–D751, 2019.

[192] B. Langmead, “Aligning short sequencing reads with bowtie,” *Current protocols in bioinformatics*, vol. 32, no. 1, pp. 11–7, 2010.

[193] R. Patro, G. Duggal, and C. Kingsford, “Salmon: accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment,” *Biorxiv*, p. 021592, 2015.

[194] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, “mirdeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades,” *Nucleic acids research*, vol. 40, no. 1, pp. 37–52, 2012.

[195] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang, and C. Lin, "e1071: Misc functions of the department of statistics (e1071), tu wien," *R package version*, vol. 1, no. 3, 2014.

[196] M. Ramos, L. Schiffer, A. Re, R. Azhar, A. Basunia, C. Rodriguez, T. Chan, P. Chapman, S. R. Davis, D. Gomez-Cabrero, *et al.*, "Software for the integration of multiomics experiments in bioconductor," *Cancer research*, vol. 77, no. 21, pp. e39–e42, 2017.

[197] H. Wickham, "testthat: Get started with testing," *The R Journal*, vol. 3, no. 1, pp. 5–10, 2011.

[198] H. Wickham, "reshape2: Flexibly reshape data: a reboot of the reshape package," *R package version*, vol. 1, no. 2, 2012.

[199] M. R. C. Team, M. R. C. Team, and S. KernSmooth, "Package grdevices," *R package version*, 2011.

[200] G. Warnes, B. Bolker, and T. Lumley, "gtools: Various r programming tools. r package version 3.5. 0," 2015.

[201] M. E. Futschik and B. Carlisle, "Noise-robust soft clustering of gene expression timecourse data," *Journal of bioinformatics and computational biology*, vol. 3, no. 04, pp. 965–988, 2005.

[202] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287, 2012.

[203] H. Wickham, L. Henry, *et al.*, "tidyr: Easily tidy data with spread ()and gather ()functions," *R package version 0.8*, vol. 2, 2018.

[204] H. Wickham, R. Francois, L. Henry, K. Müller, *et al.*, "dplyr: A grammar of data manipulation," *R package version 0.4*, vol. 3, 2015.

[205] G. Yu, "Biomedical knowledge mining using gosemsim and clusterprofiler," Oct 2020.

[206] G. R. Warnes, B. Bolker, and T. Lumley, "gtools: Various r programming tools," *R package version*, vol. 3, no. 1, 2014.

[207] H. Wickham, "stringr: Simple, consistent wrappers for common string operations," *R package version*, vol. 1, no. 0, 2017.

[208] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart," *Nature protocols*, vol. 4, no. 8, p. 1184, 2009.

[209] H. Wickham, "ggplot2," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 2, pp. 180–185, 2011.

[210] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer, 2016.

[211] W. D. Penny, "Signal processing course," *Chapter*, vol. 11, pp. 127–140, 2000.

[212] M. G. Csardi, "Package igraph," *Last accessed*, vol. 3, no. 09, p. 2013, 2013.

[213] H. Yutani, "gghighlight: Highlight lines and points inggplot2," *Manual available online at http://CRAN. R-project. org/package= gghighlight*, 2018.

[214] P. B. Ahrens, M. Solursh, and R. S. Reiter, "Stage-related capacity for limb chondrogenesis in cell culture," *Developmental biology*, vol. 60, no. 1, pp. 69–82, 1977.

[215] J. Buckwalter and H. Mankin, "Articular cartilage: tissue design and chondrocyte-matrix interactions.," *Instructional course lectures*, vol. 47, pp. 477–486, 1998.

[216] J. A. Buckwalter, H. J. Mankin, A. J. Grodzinsky, *et al.*, "Articular cartilage and osteoarthritis," *Instructional Course Lectures-American Academy of Orthopaedic Surgeons*, vol. 54, p. 465, 2005.

[217] J. Buckwalter and H. Mankin, "Articular cartilage: part i," *Journal of Bone and joint surgery*, vol. 79, no. 4, p. 600, 1997.

[218] Y. Krishnan and A. J. Grodzinsky, "Cartilage diseases," *Matrix Biology*, vol. 71, pp. 51–69, 2018.

[219] C. B. Carballo, Y. Nakagawa, I. Sekiya, and S. A. Rodeo, “Basic science of articular cartilage,” *Clinics in sports medicine*, vol. 36, no. 3, pp. 413–425, 2017.

[220] J. Yao and B. Seedhom, “Mechanical conditioning of articular cartilage to prevalent stresses,” *Rheumatology*, vol. 32, no. 11, pp. 956–965, 1993.

[221] C. Lian, X. Wang, X. Qiu, Z. Wu, B. Gao, L. Liu, G. Liang, H. Zhou, X. Yang, Y. Peng, *et al.*, “Collagen type ii suppresses articular chondrocyte hypertrophy and osteoarthritis progression by promoting integrin β1- smad1 interaction,” *Bone research*, vol. 7, no. 1, pp. 1–15, 2019.

[222] D. Eyre, “The collagens of articular cartilage,” in *Seminars in arthritis and rheumatism*, vol. 21, pp. 2–11, Elsevier, 1991.

[223] M. Van der Rest and R. Garrone, “Collagen family of proteins.,” *The FASEB journal*, vol. 5, no. 13, pp. 2814–2823, 1991.

[224] S. Wotton and V. Duance, “Type iii collagen in normal human articular cartilage,” *The Histochemical journal*, vol. 26, no. 5, pp. 412–416, 1994.

[225] B. R. Olsen, “collagen ix,” *The international journal of biochemistry & cell biology*, vol. 29, no. 4, pp. 555–558, 1997.

[226] J. A. Rada, P. K. Cornuet, and J. R. Hassell, “Regulation of corneal collagen fibrillogenesis in vitro by corneal proteoglycan (lumican and decorin) core proteins,” *Experimental eye research*, vol. 56, pp. 635–635, 1993.

[227] T. E. Hardingham and A. J. Fosang, “Proteoglycans: many forms and many functions.,” *The FASEB journal*, vol. 6, no. 3, pp. 861–870, 1992.

[228] P. J. Roughley and J. S. Mort, “The role of aggrecan in normal and osteoarthritic cartilage,” *Journal of experimental orthopaedics*, vol. 1, no. 1, pp. 1–11, 2014.

[229] J. Richmond, D. Hunter, J. Irrgang, A. M. H. Jones, L. Snyder-Mackler, M. Daniel Van Durme, C. Rubin, E. G. Matzkin, R. G. Marx, B. A. Levy, *et al.*, “The treatment of osteoarthritis (oa) of the knee,” *J Bone Joint Surg Am*, vol. 92, pp. 990–3, 2010.

[230] J. A. Buckwalter, C. Saltzman, and T. Brown, "The impact of osteoarthritis: implications for research," *Clinical Orthopaedics and Related Research®*, vol. 427, pp. S6–S15, 2004.

[231] R. C. Lawrence, D. T. Felson, C. G. Helmick, L. M. Arnold, H. Choi, R. A. Deyo, S. Gabriel, R. Hirsch, M. C. Hochberg, G. G. Hunder, *et al.*, "Estimates of the prevalence of arthritis and other rheumatic conditions in the united states: Part ii," *Arthritis & Rheumatism*, vol. 58, no. 1, pp. 26–35, 2008.

[232] H. Kotlarz, C. L. Gunnarsson, H. Fang, and J. A. Rizzo, "Osteoarthritis and absenteeism costs: evidence from us national survey data," *Journal of occupational and environmental medicine*, vol. 52, no. 3, pp. 263–268, 2010.

[233] A. Chen, C. Gupte, K. Akhtar, P. Smith, and J. Cobb, "The global economic cost of osteoarthritis: how the uk compares," *Arthritis*, vol. 2012, 2012.

[234] D. J. Hunter, D. Schofield, and E. Callander, "The individual and socioeconomic impact of osteoarthritis," *Nature Reviews Rheumatology*, vol. 10, no. 7, pp. 437–441, 2014.

[235] B. K. Hall and T. Miyake, "Divide, accumulate, differentiate: cell condensation in skeletal development revisited.," *International Journal of Developmental Biology*, vol. 39, no. 6, pp. 881–893, 2004.

[236] T. Kurth, E. Hedbom, N. Shintani, M. Sugimoto, F. Chen, M. Haspl, S. Martinovic, and E. B. Hunziker, "Chondrogenic potential of human synovial mesenchymal stem cells in alginate," *Osteoarthritis and cartilage*, vol. 15, no. 10, pp. 1178–1189, 2007.

[237] J. W. Foster, M. A. Dominguez-Steglich, S. Guioli, C. Kwok, P. A. Weller, M. Stevanović, J. Weissenbach, S. Mansour, I. D. Young, P. N. Goodfellow, *et al.*, "Campomelic dysplasia and autosomal sex reversal caused by mutations in an sry-related gene," *Nature*, vol. 372, no. 6506, pp. 525–530, 1994.

[238] T. Wagner, J. Wirth, J. Meyer, B. Zabel, M. Held, J. Zimmer, J. Pasantes, F. D. Bricarelli, J. Keutel, E. Hustert, *et al.*, "Autosomal sex reversal and campomelic

dysplasia are caused by mutations in and around the sry-related gene sox9," *Cell*, vol. 79, no. 6, pp. 1111–1120, 1994.

[239] S. Benko, J. A. Fantes, J. Amiel, D.-J. Kleinjan, S. Thomas, J. Ramsay, N. Jamshidi, A. Essafi, S. Heaney, C. T. Gordon, *et al.*, "Highly conserved non-coding elements on either side of sox9 associated with pierre robin sequence," *Nature genetics*, vol. 41, no. 3, p. 359, 2009.

[240] H. Zhao, W. Zhou, Z. Yao, Y. Wan, J. Cao, L. Zhang, J. Zhao, H. Li, R. Zhou, B. Li, *et al.*, "Foxp1/2/4 regulate endochondral ossification as a suppresser complex," *Developmental biology*, vol. 398, no. 2, pp. 242–254, 2015.

[241] P. Xu, B. Balczerski, A. Ciozda, K. Louie, V. Oralova, A. Huysseune, and J. G. Crump, "Fox proteins are modular competency factors for facial cartilage and tooth specification," *Development*, vol. 145, no. 12, p. dev165498, 2018.

[242] P. Smits, P. Li, J. Mandel, Z. Zhang, J. M. Deng, R. R. Behringer, B. De Crombrugghe, and V. Lefebvre, "The transcription factors l-sox5 and sox6 are essential for cartilage formation," *Developmental cell*, vol. 1, no. 2, pp. 277–290, 2001.

[243] T. Ikeda, S. Kamekura, A. Mabuchi, I. Kou, S. Seki, T. Takato, K. Nakamura, H. Kawaguchi, S. Ikegawa, and U.-i. Chung, "The combination of sox5, sox6, and sox9 (the sox trio) provides signals sufficient for induction of permanent cartilage," *Arthritis & Rheumatism*, vol. 50, no. 11, pp. 3561–3573, 2004.

[244] Z. Tan, B. Niu, K. Y. Tsang, I. G. Melhado, S. Ohba, X. He, Y. Huang, C. Wang, A. P. McMahon, R. Jauch, *et al.*, "Synergistic co-regulation and competition by a sox9-gli-foxa phasic transcriptional network coordinate chondrocyte differentiation transitions," *PLoS genetics*, vol. 14, no. 4, p. e1007346, 2018.

[245] F. Yano, S. Ohba, Y. Murahashi, S. Tanaka, T. Saito, and U.-i. Chung, "Runx1 contributes to articular cartilage maintenance by enhancement of cartilage matrix production and suppression of hypertrophic differentiation," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[246] D. Ikegami, H. Akiyama, A. Suzuki, T. Nakamura, T. Nakano, H. Yoshikawa, and N. Tsumaki, "Sox9 sustains chondrocyte survival and hypertrophy in part through pik3ca-akt pathways," *Development*, vol. 138, no. 8, pp. 1507–1519, 2011.

[247] J. C. Lui, S. Yue, A. Lee, B. Kikani, A. Temnycky, K. M. Barnes, and J. Baron, "Persistent sox9 expression in hypertrophic chondrocytes suppresses transdifferentiation into osteoblasts," *Bone*, vol. 125, pp. 169–177, 2019.

[248] W. Huang, X. Zhou, V. Lefebvre, and B. De Crombrugghe, "Phosphorylation of sox9 by cyclic amp-dependent protein kinase a enhances sox9's ability to transactivate acol2a1 chondrocyte-specific enhancer," *Molecular and cellular biology*, vol. 20, no. 11, pp. 4149–4158, 2000.

[249] V. Lefebvre and M. Dvir-Ginzberg, "Sox9 and the many facets of its regulation in the chondrocyte lineage," *Connective tissue research*, vol. 58, no. 1, pp. 2–14, 2017.

[250] J. A. Liu, M.-H. Wu, C. H. Yan, B. K. Chau, H. So, A. Ng, A. Chan, K. S. Cheah, J. Briscoe, and M. Cheung, "Phosphorylation of sox9 is required for neural crest delamination and is regulated downstream of bmp and canonical wnt signaling," *Proceedings of the National Academy of Sciences*, vol. 110, no. 8, pp. 2882–2887, 2013.

[251] C. Buhrmann, F. Busch, P. Shayan, and M. Shakibaei, "Sirtuin-1 (sirt1) is required for promoting chondrogenic differentiation of mesenchymal stem cells," *Journal of Biological Chemistry*, vol. 289, no. 32, pp. 22048–22062, 2014.

[252] G. Coricor and R. Serra, "Tgf-β regulates phosphorylation and stabilization of sox9 protein in chondrocytes through p38 and smad dependent mechanisms," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[253] J. C. Robins, N. Akeno, A. Mukherjee, R. R. Dalal, B. J. Aronow, P. Koopman, and T. L. Clemens, "Hypoxia induces chondrocyte-specific gene expression in mesenchymal cells in association with transcriptional activation of sox9," *Bone*, vol. 37, no. 3, pp. 313–322, 2005.

[254] E. Kozhemyakina, A. B. Lassar, and E. Zelzer, "A pathway to bone: signaling molecules and transcription factors involved in chondrocyte development and maturation," *Development*, vol. 142, no. 5, pp. 817–831, 2015.

[255] W. E. Samsa, X. Zhou, and G. Zhou, "Signaling pathways regulating cartilage growth plate formation and activity," in *Seminars in cell & developmental biology*, vol. 62, pp. 3–15, Elsevier, 2017.

[256] V. Lefebvre, M. Angelozzi, and A. Haseeb, "Sox9 in cartilage development and disease," *Current opinion in cell biology*, vol. 61, pp. 39–47, 2019.

[257] A. Woods, G. Wang, and F. Beier, "Rhoa/rock signaling regulates sox9 expression and actin organization during chondrogenesis," *Journal of Biological Chemistry*, vol. 280, no. 12, pp. 11626–11634, 2005.

[258] M. M.-G. Sun and F. Beier, "Chondrocyte hypertrophy in skeletal development, growth, and disease," *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 102, no. 1, pp. 74–82, 2014.

[259] P. Singh, K. B. Marcu, M. B. Goldring, and M. Otero, "Phenotypic instability of chondrocytes in osteoarthritis: on a path to hypertrophy," *Annals of the New York Academy of Sciences*, vol. 1442, no. 1, pp. 17–34, 2019.

[260] V. Lefebvre, R. R. Behringer, and B. De Crombrugghe, "L-sox5, sox6 and sox9 control essential steps of the chondrocyte differentiation pathway," *Osteoarthritis and Cartilage*, vol. 9, pp. S69–S75, 2001.

[261] H. Kishimoto, M. Akagi, S. Zushi, T. Teramura, Y. Onodera, T. Sawamura, and C. Hamanishi, "Induction of hypertrophic chondrocyte-like phenotypes by oxidized ldl in cultured bovine articular chondrocytes through increase in oxidative stress," *Osteoarthritis and cartilage*, vol. 18, no. 10, pp. 1284–1290, 2010.

[262] M. D'Angelo, Z. Yan, M. Nooreyazdan, M. Pacifici, D. Sarment, P. C. Billings, and P. S. Leboy, "Mmp-13 is induced during chondrocyte hypertrophy," *Journal of cellular biochemistry*, vol. 77, no. 4, pp. 678–693, 2000.

[263] S. E. Usmani, M. A. Pest, G. Kim, S. N. Ohora, L. Qin, and F. Beier, "Transforming growth factor alpha controls the transition from hypertrophic cartilage to bone during endochondral bone growth," *Bone*, vol. 51, no. 1, pp. 131–141, 2012.

[264] E. Zelzer, R. Mamluk, N. Ferrara, R. S. Johnson, E. Schipani, and B. R. Olsen, "Vegfa is necessary for chondrocyte survival during bone development," *Development*, vol. 131, no. 9, pp. 2161–2171, 2004.

[265] R. B. Vega, K. Matsuda, J. Oh, A. C. Barbosa, X. Yang, E. Meadows, J. McAnally, C. Pomajzl, J. M. Shelton, J. A. Richardson, *et al.*, "Histone deacetylase 4 controls chondrocyte hypertrophy during skeletogenesis," *Cell*, vol. 119, no. 4, pp. 555–566, 2004.

[266] M. Caron, P. Emans, A. Cremers, D. Surtel, M. Coolsen, L. Van Rhijn, and T. Welting, "Hypertrophic differentiation during chondrogenic differentiation of progenitor cells is stimulated by bmp-2 but suppressed by bmp-7," *Osteoarthritis and Cartilage*, vol. 21, no. 4, pp. 604–613, 2013.

[267] C. Wu, B. Tian, X. Qu, F. Liu, T. Tang, A. Qin, Z. Zhu, and K. Dai, "Micrornas play a role in chondrogenesis and osteoarthritis," *International journal of molecular medicine*, vol. 34, no. 1, pp. 13–23, 2014.

[268] T. Kobayashi, J. Lu, B. S. Cobb, S. J. Rodda, A. P. McMahon, E. Schipani, M. Merkenschlager, and H. M. Kronenberg, "Dicer-dependent pathways regulate chondrocyte proliferation and differentiation," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 1949–1954, 2008.

[269] Z. Zhang, J. R. O'Rourke, M. T. McManus, M. Lewandoski, B. D. Harfe, and X. Sun, "The microrna-processing enzyme dicer is dispensable for somite segmentation but essential for limb bud positioning," *Developmental biology*, vol. 351, no. 2, pp. 254–265, 2011.

[270] Z.-j. Liang, H. Zhuang, G.-x. Wang, Z. Li, H.-t. Zhang, T.-q. Yu, and B.-d. Zhang, "Mirna-140 is a negative feedback regulator of mmp-13 in il-1β-stimulated human articular chondrocyte c28/i2 cells," *Inflammation Research*, vol. 61, no. 5, pp. 503–509, 2012.

[271] L. Tuddenham, G. Wheeler, S. Ntounia-Fousara, J. Waters, M. K. Hajihosseini, I. Clark, and T. Dalmay, "The cartilage specific microrna-140 targets histone deacetylase 4 in mouse cells," *FEBS letters*, vol. 580, no. 17, pp. 4214–4217, 2006.

[272] J. Lu, Y. Sun, Q. Ge, H. Teng, and Q. Jiang, "Histone deacetylase 4 alters cartilage homeostasis in human osteoarthritis," *BMC musculoskeletal disorders*, vol. 15, no. 1, p. 438, 2014.

[273] F. E. Nicolas, H. Pais, F. Schwach, M. Lindow, S. Kauppinen, V. Moulton, and T. Dalmay, "mrna expression profiling reveals conserved and non-conserved mir-140 targets," *RNA biology*, vol. 8, no. 4, pp. 607–615, 2011.

[274] S. Yamashita, S. Miyaki, Y. Kato, S. Yokoyama, T. Sato, F. Barrionuevo, H. Akiyama, G. Scherer, S. Takada, and H. Asahara, "L-sox5 and sox6 proteins enhance chondrogenic mir-140 microrna expression by strengthening dimeric sox9 activity," *Journal of Biological Chemistry*, vol. 287, no. 26, pp. 22206–22215, 2012.

[275] D. Guérit, D. Philipot, J.-M. Brondello, P. Chuchana, C. Jorgensen, and D. Noël, "Inhibitory effect of mir-29a on the chondrogenic differentiation of mesenchymal stem cells," *Osteoarthritis and Cartilage*, vol. 20, p. S52, 2012.

[276] L. T. Le, T. E. Swingler, N. Crowe, T. L. Vincent, M. J. Barter, S. T. Donell, A. M. Delany, T. Dalmay, D. A. Young, and I. M. Clark, "The microrna-29 family in cartilage homeostasis and osteoarthritis," *Journal of molecular medicine*, vol. 94, no. 5, pp. 583–596, 2016.

[277] L. Dai, X. Zhang, X. Hu, C. Zhou, and Y. Ao, "Silencing of microrna-101 prevents il-1β-induced extracellular matrix degradation in chondrocytes," *Arthritis research & therapy*, vol. 14, no. 6, p. R268, 2012.

[278] A. Martinez-Sanchez, K. A. Dudek, and C. L. Murphy, "Regulation of human chondrocyte function through direct inhibition of cartilage master regulator sox9 by microrna-145 (mirna-145)," *Journal of Biological Chemistry*, vol. 287, no. 2, pp. 916–924, 2012.

[279] S. Lee, D. S. Yoon, S. Paik, K.-M. Lee, Y. Jang, and J. W. Lee, "microrna-495 inhibits chondrogenic differentiation in human mesenchymal stem cells by targeting sox9," *Stem cells and development*, vol. 23, no. 15, pp. 1798–1808, 2014.

[280] T. Ukai, M. Sato, H. Akutsu, A. Umezawa, and J. Mochida, "Microrna-199a-3p, microrna-193b, and microrna-320c are correlated to aging and regulate human cartilage metabolism," *Journal of Orthopaedic Research*, vol. 30, no. 12, pp. 1915–1922, 2012.

[281] T. Matsukawa, T. Sakai, T. Yonezawa, H. Hiraiwa, T. Hamada, M. Nakashima, Y. Ono, S. Ishizuka, H. Nakahara, M. K. Lotz, *et al.*, "Microrna-125b regulates the expression of aggrecanase-1 (adamts-4) in human osteoarthritic chondrocytes," *Arthritis research & therapy*, vol. 15, no. 1, p. R28, 2013.

[282] N. Akhtar, Z. Rasheed, S. Ramamurthy, A. N. Anbazhagan, F. R. Voss, and T. M. Haqqi, "Microrna-27b regulates the expression of matrix metalloproteinase 13 in human osteoarthritis chondrocytes," *Arthritis & Rheumatism*, vol. 62, no. 5, pp. 1361–1371, 2010.

[283] M. J. Moore, T. K. Scheel, J. M. Luna, C. Y. Park, J. J. Fak, E. Nishiuchi, C. M. Rice, and R. B. Darnell, "mirna–target chimeras reveal mirna 3-end pairing as a major determinant of argonaute target specificity," *Nature communications*, vol. 6, no. 1, pp. 1–17, 2015.

[284] R. Akhter, Y. Shao, M. Shaw, S. Formica, M. Khrestian, J. B. Leverenz, and L. M. Bekris, "Regulation of adam10 by mir-140-5p and potential relevance for alzheimer's disease," *Neurobiology of aging*, vol. 63, pp. 110–119, 2018.

[285] A. D. Murdoch, L. M. Grady, M. P. Ablett, T. Katopodi, R. S. Meadows, and T. E. Hardingham, "Chondrogenic differentiation of human bone marrow stem cells in transwell cultures: Generation of scaffold-free cartilage," *Stem cells*, vol. 25, no. 11, pp. 2786–2796, 2007.

[286] P. Du, W. A. Kibbe, and S. M. Lin, "lumi: a pipeline for processing illumina microarray," *Bioinformatics*, vol. 24, no. 13, pp. 1547–1548, 2008.

[287] S. Gubian, A. Sewer, and P. SA, “Description of eximir,” 2013.

[288] J. Soul, T. E. Hardingham, R. P. Boot-Handford, and J.-M. Schwartz, “Skeletalvis: an exploration and meta-analysis data portal of cross-species skeletal transcriptomics data,” *Bioinformatics*, vol. 35, no. 13, pp. 2283–2290, 2019.

[289] N. P. Huynh, B. Zhang, and F. Guilak, “High-depth transcriptomic profiling reveals the temporal gene signature of human mesenchymal stem cells during chondrogenesis,” *The FASEB Journal*, vol. 33, no. 1, pp. 358–372, 2019.

[290] A. H. Huang, A. Stein, and R. L. Mauck, “Evaluation of the complex transcriptional topography of mesenchymal stem cell chondrogenesis for cartilage tissue engineering,” *Tissue Engineering Part A*, vol. 16, no. 9, pp. 2699–2708, 2010.

[291] C. L. L. Cardenas, I. S. Henaoui, E. Courcot, C. Roderburg, C. Cauffiez, S. Aubert, M.-C. Copin, B. Wallaert, F. Glowacki, E. Dewaeles, *et al.*, “mir-199a-5p is upregulated during fibrogenic response to tissue injury and mediates tgfbeta-induced lung fibroblast activation by targeting caveolin-1,” *PLoS Genet*, vol. 9, no. 2, p. e1003291, 2013.

[292] P.-x. Zhang, J. Cheng, S. Zou, A. D. D’Souza, J. L. Koff, J. Lu, P. J. Lee, D. S. Krause, M. E. Egan, and E. M. Bruscia, “Pharmacological modulation of the akt/microrna-199a-5p/cav1 pathway ameliorates cystic fibrosis lung hyper-inflammation,” *Nature communications*, vol. 6, no. 1, pp. 1–13, 2015.

[293] J. Wang, M.-y. Chen, J.-f. Chen, Q.-l. Ren, J.-q. Zhang, H. Cao, B.-s. Xing, and C.-y. Pan, “Lncrna imflnc1 promotes porcine intramuscular adipocyte adipogenesis by sponging mir-199a-5p to up-regulate cav-1,” *BMC molecular and cell biology*, vol. 21, no. 1, pp. 1–16, 2020.

[294] F. Du, Y. Zhang, Q. Xu, Y. Teng, M. Tao, A. F. Chen, and R. Jiang, “Preeclampsia serum increases cav1 expression and cell permeability of human renal glomerular endothelial cells via down-regulating mir-199a-5p, mir-199b-5p, mir-204,” *Placenta*, vol. 99, pp. 141–151, 2020.

[295] M. Carlson, ". lumihumanall.db: Illumina human illumina expression annotation data (chip lumihumanall)," *R*, vol. 1.22.0, 2013.

[296] C. Dubroca, X. Loyer, K. Retailleau, G. Loirand, P. Pacaud, O. Feron, J.-L. Balligand, B. I. Lévy, C. Heymes, and D. Henrion, "Rhoa activation and interaction with caveolin-1 are critical for pressure-induced myogenic tone in rat mesenteric resistance arteries," *Cardiovascular research*, vol. 73, no. 1, pp. 190–197, 2007.

[297] F. Peng, B. Zhang, D. Wu, A. J. Ingram, B. Gao, and J. C. Krepinsky, "Tgfβ-induced rhoa activation and fibronectin production in mesangial cells require caveolae," *American Journal of Physiology-Renal Physiology*, vol. 295, no. 1, pp. F153–F164, 2008.

[298] B. Joshi, S. S. Strugnell, J. G. Goetz, L. D. Kojic, M. E. Cox, O. L. Griffith, S. K. Chan, S. J. Jones, S.-P. Leung, H. Masoudi, *et al.*, "Phosphorylated caveolin-1 regulates rho/rock-dependent focal adhesion dynamics and tumor cell migration and invasion," *Cancer research*, vol. 68, no. 20, pp. 8210–8220, 2008.

[299] F. M. Vega and A. J. Ridley, "Rho gtpases in cancer cell biology," *FEBS letters*, vol. 582, no. 14, pp. 2093–2101, 2008.

[300] S. J. Heasman and A. J. Ridley, "Mammalian rho gtpases: new insights into their functions from in vivo studies," *Nature reviews Molecular cell biology*, vol. 9, no. 9, pp. 690–701, 2008.

[301] S. Arber, F. A. Barbayannis, H. Hanser, C. Schneider, C. A. Stanyon, O. Bernard, and P. Caroni, "Regulation of actin dynamics through phosphorylation of cofilin by lim-kinase," *Nature*, vol. 393, no. 6687, pp. 805–809, 1998.

[302] M. Maekawa, T. Ishizaki, S. Boku, N. Watanabe, A. Fujita, A. Iwamatsu, T. Obinata, K. Ohashi, K. Mizuno, and S. Narumiya, "Signaling from rho to the actin cytoskeleton through protein kinases rock and lim-kinase," *Science*, vol. 285, no. 5429, pp. 895–898, 1999.

[303] F. Beier and R. F. Loeser, "Biology and pathology of rho gtpase, pi-3 kinase-akt, and

map kinase signaling pathways in chondrocytes," *Journal of cellular biochemistry*, vol. 110, no. 3, pp. 573–580, 2010.

[304] A. Woods and F. Beier, "Rhoa/rock signaling regulates chondrogenesis in a context-dependent manner," *Journal of Biological Chemistry*, vol. 281, no. 19, pp. 13134–13140, 2006.

[305] S. R. Tew and T. E. Hardingham, "Regulation of sox9 mrna in human articular chondrocytes involving p38 mapk activation and mrna stabilization," *Journal of Biological Chemistry*, vol. 281, no. 51, pp. 39471–39479, 2006.

[306] T. Xu, M. Wu, J. Feng, X. Lin, and Z. Gu, "Rhoa/rho kinase signaling regulates transforming growth factor-β1-induced chondrogenesis and actin organization of synovium-derived mesenchymal stem cells through interaction with the smad pathway," *International journal of molecular medicine*, vol. 30, no. 5, pp. 1119–1125, 2012.

[307] V. Lefebvre, P. Li, and B. De Crombrugghe, "A new long form of sox5 (l-sox5), sox6 and sox9 are coexpressed in chondrogenesis and cooperatively activate the type ii collagen gene," *The EMBO journal*, vol. 17, no. 19, pp. 5718–5733, 1998.

[308] Y. Han and V. Lefebvre, "L-sox5 and sox6 drive expression of the aggrecan gene in cartilage by securing binding of sox9 to a far-upstream enhancer," *Molecular and cellular biology*, vol. 28, no. 16, pp. 4999–5013, 2008.

[309] D. Kumar and A. B. Lassar, "The transcriptional activity of sox9 in chondrocytes is regulated by rhoa signaling and actin polymerization," *Molecular and cellular biology*, vol. 29, no. 15, pp. 4262–4273, 2009.

[310] D. R. Haudenschild, J. Chen, N. Pang, M. K. Lotz, and D. D. D'Lima, "Rho kinase–dependent activation of sox9 in chondrocytes," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 62, no. 1, pp. 191–200, 2010.

[311] S. Preiss, A. Argentaro, A. Clayton, A. John, D. A. Jans, T. Ogata, T. Nagai, I. Barroso, A. J. Schafer, and V. R. Harley, "Compound effects of point mutations causing

campomelic dysplasia/autosomal sex reversal upon sox9 structure, nuclear transport, dna binding, and transcriptional activation," *Journal of Biological Chemistry*, vol. 276, no. 30, pp. 27864–27872, 2001.

[312] S. Malki, S. Nef, C. Notarnicola, L. Thevenet, S. Gasca, C. Mejean, P. Berta, F. Poulat, and B. Boizet-Bonhoure, "Prostaglandin d2 induces nuclear import of the sex-determining factor sox9 via its camp-pka phosphorylation," *The EMBO journal*, vol. 24, no. 10, pp. 1798–1809, 2005.

[313] H. Chikuda, F. Kugimiya, K. Hoshi, T. Ikeda, T. Ogasawara, T. Shimoaka, H. Kawano, S. Kamekura, A. Tsuchida, N. Yokoi, *et al.*, "Cyclic gmp-dependent protein kinase ii is a molecular switch from proliferation to hypertrophic differentiation of chondrocytes," *Genes & Development*, vol. 18, no. 19, pp. 2418–2429, 2004.

[314] T. Furumatsu, M. Tsuda, N. Taniguchi, Y. Tajima, and H. Asahara, "Smad3 induces chondrogenesis through the activation of sox9 via creb-binding protein/p300 recruitment," *Journal of Biological Chemistry*, vol. 280, no. 9, pp. 8343–8350, 2005.

[315] A. K. Kamaraju and A. B. Roberts, "Role of rho/rock and p38 map kinase pathways in transforming growth factor-β-mediated smad-dependent growth inhibition of human breast carcinoma cells in vivo," *Journal of Biological Chemistry*, vol. 280, no. 2, pp. 1024–1036, 2005.

[316] A. J. Hollins, L. Campbell, M. Gumbleton, and D. J. Evans, "Caveolin expression during chondrogenesis in the avian limb," *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, vol. 225, no. 2, pp. 205–211, 2002.

[317] J. Rubin, Z. Schwartz, B. D. Boyan, X. Fan, N. Case, B. Sen, M. Drab, D. Smith, M. Aleman, K. L. Wong, *et al.*, "Caveolin-1 knockout mice have increased bone size and stiffness," *Journal of Bone and Mineral Research*, vol. 22, no. 9, pp. 1408–1418, 2007.

[318] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, "Celldesigner 3.5: a versatile modeling tool for biochemical networks," *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.

[319] R. Mishra, L. Zhu, R. L. Eckert, and M. S. Simonson, "Tgf-β-regulated collagen type i accumulation: role of src-based signals," *American Journal of Physiology-Cell Physiology*, vol. 292, no. 4, pp. C1361–C1369, 2007.

[320] D. Pala, M. Kapoor, A. Woods, L. Kennedy, S. Liu, S. Chen, L. Bursell, K. M. Lyons, D. E. Carter, F. Beier, *et al.*, "Focal adhesion kinase/src suppresses early chondrogenesis: central role of ccn2," *Journal of Biological Chemistry*, vol. 283, no. 14, pp. 9239–9247, 2008.

[321] S.-M. Dai, Z.-Z. Shan, H. Nakamura, K. Masuko-Hongo, T. Kato, K. Nishioka, and K. Yudoh, "Catabolic stress induces features of chondrocyte senescence through overexpression of caveolin 1: Possible involvement of caveolin 1–induced down-regulation of articular chondrocytes in the pathogenesis of osteoarthritis," *Arthritis & Rheumatism*, vol. 54, no. 3, pp. 818–831, 2006.

[322] W. Peng, D. He, B. Shan, J. Wang, W. Shi, W. Zhao, Z. Peng, Q. Luo, M. Duan, B. Li, *et al.*, "Linc81507 act as a competing endogenous rna of mir-199b-5p to facilitate nsclc proliferation and metastasis via regulating the cav1/stat3 pathway," *Cell death & disease*, vol. 10, no. 7, pp. 1–15, 2019.

[323] G. Wang, A. Woods, S. Sabari, L. Pagnotta, L.-A. Stanton, and F. Beier, "Rhoa/rock signaling suppresses hypertrophic chondrocyte differentiation," *Journal of Biological Chemistry*, vol. 279, no. 13, pp. 13205–13214, 2004.

[324] L. Garzia, I. Andolfo, E. Cusanelli, N. Marino, G. Petrosino, D. De Martino, V. Esposito, A. Galeone, L. Navas, S. Esposito, *et al.*, "Microrna-199b-5p impairs cancer stem cells through negative regulation of hes1 in medulloblastoma," *PloS one*, vol. 4, no. 3, p. e4998, 2009.

[325] X. Zhang, C. Wei, J. Li, J. Liu, and J. Qu, "Microrna-361-5p inhibits epithelial-to-mesenchymal transition of glioma cells through targeting twist1," *Oncology reports*, vol. 37, no. 3, pp. 1849–1856, 2017.

[326] A. Kanitz, J. Imig, P. J. Dziunycz, A. Primorac, A. Galgano, G. F. Hofbauer, A. P. Gerber, and M. Detmar, "The expression levels of microrna-361-5p and its target

vegfa are inversely correlated in human cutaneous squamous cell carcinoma," *PLoS One*, vol. 7, no. 11, p. e49568, 2012.

[327] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, M. Hucka, G. Jalowicki, S. Keating, V. Knight-Schrijver, *et al.*, "Biomodels: ten-year anniversary," *Nucleic acids research*, vol. 43, no. D1, pp. D542–D548, 2015.

[328] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information processing letters*, vol. 85, no. 6, pp. 317–325, 2003.

[329] I. Moser, "Hooke-jeeves revisited," in *2009 IEEE Congress on Evolutionary Computation*, pp. 2670–2676, IEEE, 2009.

[330] J. Sun, J. M. Garibaldi, and C. Hodgman, "Parameter estimation using metaheuristics in systems biology: a comprehensive review," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 9, no. 1, pp. 185–202, 2011.

[331] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter estimation in biochemical pathways: a comparison of global optimization methods," *Genome research*, vol. 13, no. 11, pp. 2467–2474, 2003.

[332] S. Fleige, V. Walf, S. Huch, C. Prgomet, J. Sehm, and M. W. Pfaffl, "Comparison of relative mrna quantification models and the impact of rna integrity in quantitative real-time rt-pcr," *Biotechnology letters*, vol. 28, no. 19, pp. 1601–1613, 2006.

[333] J.-H. S. Hofmeyr and H. Cornish-Bowden, "The reversible hill equation: how to incorporate cooperative enzymes into metabolic models," *Bioinformatics*, vol. 13, no. 4, pp. 377–385, 1997.

[334] C. Balcells, I. Pastor, E. Vilaseca, S. Madurga, M. Cascante, and F. Mas, "Macromolecular crowding effect upon in vitro enzyme kinetics: mixed activation–diffusion control of the oxidation of nadh by pyruvate catalyzed by lactate dehydrogenase," *The Journal of Physical Chemistry B*, vol. 118, no. 15, pp. 4062–4068, 2014.

[335] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[336] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, vol. 50, pp. 71–91, 2019.

[337] C. Outeiral, M. Strahm, J. Shi, G. M. Morris, S. C. Benjamin, and C. M. Deane, "The prospects of quantum computing in computational molecular biology," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 11, no. 1, p. e1481, 2021.

[338] C. Casella, I. Lipp, A. Rosser, D. K. Jones, and C. Metzler-Baddeley, "A critical review of white matter changes in huntingtons disease," *Movement Disorders*, vol. 35, no. 8, pp. 1302–1311, 2020.

[339] H. Rosas, A. Liu, S. Hersch, M. Glessner, R. Ferrante, D. Salat, A. van Der Kouwe, B. Jenkins, A. Dale, and B. Fischl, "Regional and progressive thinning of the cortical ribbon in huntingtons disease," *Neurology*, vol. 58, no. 5, pp. 695–701, 2002.

[340] G. P. Bates, R. Dorsey, J. F. Gusella, M. R. Hayden, C. Kay, B. R. Leavitt, M. Nance, C. A. Ross, R. I. Scahill, R. Wetzel, *et al.*, "Huntington disease," *Nature reviews Disease primers*, vol. 1, no. 1, pp. 1–21, 2015.

[341] A. Semaka, J. Collins, and M. Hayden, "Unstable familial transmissions of huntington disease alleles with 27–35 cag repeats (intermediate alleles)," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 153, no. 1, pp. 314–320, 2010.

[342] A. Rosenblatt, B. V. Kumar, A. Mo, C. S. Welsh, R. L. Margolis, and C. A. Ross, "Age, cag repeat length, and clinical progression in huntington's disease," *Movement disorders*, vol. 27, no. 2, pp. 272–276, 2012.

[343] J. Achenbach, C. Thiels, T. Lücke, and C. Saft, "Clinical manifestation of juvenile and pediatric hd patients: a retrospective case series," *Brain Sciences*, vol. 10, no. 6, p. 340, 2020.

[344] M. D. Rawlins, N. S. Wexler, A. R. Wexler, S. J. Tabrizi, I. Douglas, S. J. Evans, and

L. Smeeth, “The prevalence of huntington’s disease,” *Neuroepidemiology*, vol. 46, no. 2, pp. 144–153, 2016.

[345] S. J. Evans, I. Douglas, M. D. Rawlins, N. S. Wexler, S. J. Tabrizi, and L. Smeeth, “Prevalence of adult huntington’s disease in the uk based on diagnoses recorded in general practice records,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 84, no. 10, pp. 1156–1160, 2013.

[346] C. Ohlmeier, K.-U. Saum, W. Galetzka, D. Beier, and H. Gothe, “Epidemiology and health care utilization of patients suffering from huntingtons disease in germany: real world evidence based on german claims data,” *BMC neurology*, vol. 19, no. 1, pp. 1–8, 2019.

[347] N. Georgiou, J. L. Bradshaw, E. Chiu, A. Tudor, L. O’Gorman, and J. G. Phillips, “Differential clinical and motor control function in a pair of monozygotic twins with huntington’s disease,” *Movement disorders: official journal of the Movement Disorder Society*, vol. 14, no. 2, pp. 320–325, 1999.

[348] F. R. Fusco, Q. Chen, W. J. Lamoreaux, G. Figueredo-Cardenas, Y. Jiao, J. A. Coffman, D. J. Surmeier, M. G. Honig, L. R. Carlock, and A. Reiner, “Cellular localization of huntingtin in striatal and cortical neurons in rats: lack of correlation with neuronal vulnerability in huntingtons disease,” *Journal of Neuroscience*, vol. 19, no. 4, pp. 1189–1202, 1999.

[349] H. Takano and J. F. Gusella, “The predominantly heat-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an nf-kb/rel/dorsal family transcription factor,” *BMC neuroscience*, vol. 3, no. 1, pp. 1–13, 2002.

[350] J. Cornett, F. Cao, C.-E. Wang, C. A. Ross, G. P. Bates, S.-H. Li, and X.-J. Li, “Polyglutamine expansion of huntingtin impairs its nuclear export,” *Nature genetics*, vol. 37, no. 2, pp. 198–204, 2005.

[351] G. Hoffner, P. Kahlem, and P. Djian, “Perinuclear localization of huntingtin as a consequence of its binding to microtubules through an interaction with β-tubulin: relevance to huntington’s disease,” *Journal of cell science*, vol. 115, no. 5, pp. 941–948, 2002.

[352] J. P. Caviston, J. L. Ross, S. M. Antony, M. Tokito, and E. L. Holzbaur, "Huntingtin facilitates dynein/dynactin-mediated vesicle transport," *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 10045–10050, 2007.

[353] C. Zuccato, M. Tartari, A. Crotti, D. Goffredo, M. Valenza, L. Conti, T. Cataudella, B. R. Leavitt, M. R. Hayden, T. Timmusk, *et al.*, "Huntingtin interacts with rest/nrsf to modulate the transcription of nrse-controlled neuronal genes," *Nature genetics*, vol. 35, no. 1, pp. 76–83, 2003.

[354] K. N. McFarland, M. N. Huizenga, S. B. Darnell, G. R. Sangrey, O. Berezovska, J.-H. J. Cha, T. F. Outeiro, and G. Sadri-Vakili, "Mecp2: a novel huntingtin interactor," *Human Molecular Genetics*, vol. 23, no. 4, pp. 1036–1044, 2014.

[355] M. Jimenez-Sanchez, F. Licitra, B. R. Underwood, and D. C. Rubinsztein, "Huntingtons disease: mechanisms of pathogenesis and therapeutic strategies," *Cold Spring Harbor perspectives in medicine*, vol. 7, no. 7, p. a024240, 2017.

[356] A. N. Packer, Y. Xing, S. Q. Harper, L. Jones, and B. L. Davidson, "The bifunctional microrna mir-9/mir-9* regulates rest and corest and is downregulated in huntington's disease," *Journal of Neuroscience*, vol. 28, no. 53, pp. 14341–14346, 2008.

[357] R. Johnson and N. J. Buckley, "Gene dysregulation in huntingtons disease: Rest, micrornas and beyond," *Neuromolecular medicine*, vol. 11, no. 3, pp. 183–199, 2009.

[358] M. E. Andrés, C. Burger, M. J. Peral-Rubio, E. Battaglioli, M. E. Anderson, J. Grimes, J. Dallman, N. Ballas, and G. Mandel, "Corest: a functional corepressor required for regulation of neural-specific gene expression," *Proceedings of the National Academy of Sciences*, vol. 96, no. 17, pp. 9873–9878, 1999.

[359] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS genetics*, vol. 3, no. 9, p. e161, 2007.

[360] S. Quinlan, A. Kenny, M. Medina, T. Engel, and E. M. Jimenez-Mateos, "Micrornas in neurodegenerative diseases," *International review of cell and molecular biology*, vol. 334, pp. 309–343, 2017.

[361] P. M. Gaughwin, M. Ciesla, N. Lahiri, S. J. Tabrizi, P. Brundin, and M. Björkqvist, "Hsa-mir-34b is a plasma-stable microrna that is elevated in pre-manifest huntington's disease," *Human molecular genetics*, vol. 20, no. 11, pp. 2225–2237, 2011.

[362] E. R. Reed, J. C. Latourelle, J. H. Bockholt, J. Bregu, J. Smock, J. S. Paulsen, R. H. Myers, P.-H. C. ancillary study investigators, *et al.*, "Micrornas in csf as prodromal biomarkers for huntington disease in the predict-hd study," *Neurology*, vol. 90, no. 4, pp. e264–e272, 2018.

[363] L. Wang and L. Zhang, "Circulating exosomal mirna as diagnostic biomarkers of neurodegenerative diseases," *Frontiers in molecular neuroscience*, vol. 13, p. 53, 2020.

[364] R. Sager, "Big data to good data: Andrew ng urges ml community to be more data-centric and less model-centric," 2021.

[365] G. Bartzokis, P. H. Lu, T. A. Tishler, S. M. Fong, B. Oluwadara, J. P. Finn, D. Huang, Y. Bordelon, J. Mintz, and S. Perlman, "Myelin breakdown and iron changes in huntingtons disease: pathogenesis and treatment implications," *Neurochemical research*, vol. 32, no. 10, pp. 1655–1664, 2007.

[366] J. M. Dietschy, "Central nervous system: cholesterol turnover, brain development and neurodegeneration," 2009.

[367] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[368] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[369] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[370] K. Muralidhar, "The right way of using smote with cross-validation," 2021.

[371] C. Soneson, M. I. Love, and M. D. Robinson, "Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences," *F1000Research*, vol. 4, 2015.

[372] A. A. Millar and P. M. Waterhouse, "Plant and animal micrornas: similarities and differences," *Functional & integrative genomics*, vol. 5, no. 3, pp. 129–135, 2005.

[373] A. Bohler, L. M. Eijssen, M. P. van Iersel, C. Leemans, E. L. Willighagen, M. Kutmon, M. Jaillard, and C. T. Evelo, "Automatically visualise and analyse data on pathways using pathvisiorpc from any programming environment," *BMC bioinformatics*, vol. 16, no. 1, p. 267, 2015.

[374] T. Swingler, L. Niu, P. Smith, P. Paddy, L. Le, M. Barter, D. Young, and I. Clark, "The function of micrornas in cartilage and osteoarthritis," *Clinical and experimental rheumatology*, vol. 37, no. 5, pp. 40–47, 2019.

[375] V. L. LaPointe, A. Verpoorte, and M. M. Stevens, "The changing integrin expression and a role for integrin β8 in the chondrogenic differentiation of mesenchymal stem cells," *PLoS One*, vol. 8, no. 11, p. e82035, 2013.

[376] E. Danen, P. Sonneveld, C. Brakebusch, R. Fassler, and A. Sonnenberg, "The fibronectin-binding integrins 51 and v3 differentially modulate rhoa-gtp loading, organization of cell matrix adhesions, and fibronectin fibrillogenesis," *J. Cell Biol*, vol. 159, pp. 1071–1086, 2002.

[377] G. Corda and A. Sala, "Non-canonical wnt/pcp signalling in cancer: Fzd6 takes centre stage," *Oncogenesis*, vol. 6, no. 7, pp. e364–e364, 2017.

CHAPTER 8

APPENDIX

8.1 Appendix - A. Publications

One high-impact publication so far. Two further publications are aimed to be submitted by the end of 2021.

Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Data and Text Mining

TimiRGeN: R/Bioconductor package for time series microRNA-mRNA integration and analysis

Patel K [1], Chandrasegaran S [1] Clark IM [2], Proctor CJ [1], Young DA [3], Shanley DP* [1]

[1] Campus for Ageing and Vitality, Biosciences Institute, Newcastle University, Newcastle upon-Tyne, NE4 5PL, UK.
[2] School of Biological Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.
[3] Life Science Centre, Biosciences Institute, Newcastle University, Newcastle upon-Tyne, NE1 4EP, UK.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The analysis of longitudinal datasets and construction of gene regulatory networks provide a valuable means to disentangle the complexity of microRNA-mRNA interactions. However, there are no computational tools that can integrate, conduct functional analysis and generate detailed networks from longitudinal microRNA-mRNA datasets.
Results: We present *TimiRGeN*, an *R* package that uses time point based differential expression results to identify miRNA-mRNA interactions influencing signalling pathways of interest. miRNA-mRNA interactions can be visualised in *R* or exported to *PathVisio* or *Cytoscape*. The output can be used for hypothesis generation and directing *in vitro* or further *in silico* work such as gene regulatory network construction.
Availability and implementation: *TimiRGeN* is available for download on Bioconductor (https://bioconductor.org/packages/TimiRGeN) and requires *R* v4.0.2 or newer and *BiocManager* v3.12 or newer.
Contact: k.patel5@ncl.ac.uk, daryl.shanley@ncl.ac.uk
Supplementary information: Supplementary data is available at *Bioinformatics* online.

1 Introduction

microRNAs (miRNAs) are single-stranded functional RNAs, around 16-22 nucleotides long which target specific mRNAs for degradation or translational repression; thus affecting protein levels (Selbach *et al*., 2008). Targeting is achieved by complementary binding between the 3'UTR of the target mRNA and a 7-8 nucleotide sequence found on the 5'UTR of the miRNA, known as the seed sequence (Bartel., 2004). There is increased clinical interest in miRNAs for several reasons: 1) miRNAs can be tested in animal models to understand human diseases and conditions. An example is miR-140-5p which is up-regulated during chondrogenesis and down-regulated during osteoarthritis (Barter *et al*., 2015; Miyaki *et al*., 2010). 2) miRNAs can be secreted via exosomes into surrounding blood, extracellular matrix and urine (Leidinger *et al*., 2013; Chaturvedi *et al*., 2015; Chen *et al*., 2017). Their presence in body fluids provides valuable non-invasive biomarkers to assess the state of difficult to access tissues such as tumours, brain and bone. 3) Lastly, miRNAs have potential as therapeutic agents as they modulate expression of specific mRNAs (Schwarzenbach *et al*., 2014).

However, in the laboratory, miRNAs are difficult to study, primarily because a single miRNA can regulate many mRNAs and a single mRNA can be regulated by multiple miRNAs. miRNA-mRNA interactome studies report over 18,000 interactions in HEK293 cells and over 34,000 interactions in human hepatoma cells (Helwak *et al*., 2013; Moore *et al*., 2015). A complementary strategy is to use a computational approach. The analysis of longitudinal miRNA-mRNA expression data, construction of Gene Regulatory Networks (GRNs) and subsequent dynamic modelling, is a particularly useful means to gain a better understanding of miRNA-mRNA interactions (Qin *et al*., 2015; Proctor *et al*., 2017; Ooi *et al*., 2018). GRNs are useful tools for integrating mutli-omic data on mechanistic schematics. Yet, currently there is no computational tool that can handle longitudinal miRNA-mRNA datasets and reduce the volume of data to an extent where GRN construction is possible, and this is presented in Table 1.

Tool name	Availability	Time	Funct analysis	Reduction	Updated
anamiR	Bioc	×	✓:Kegg,React,+	✓	2018
DREM2	Install	✓	✓:GO	×	2020
MAGIA2	Online	×	✓:DAVID	✓	2012
miARMa-seq	Install	✓	✓:GO,Kegg	×	2019
miRComb	SF	✓	✓:GO,Kegg	✓	2020
miRIntegrator	Bioc	×	✓:Kegg,React	✓	2016
miRNet	Online	×	✓:GO,Kegg	×	2021
miRTarVis+	Online	×	×	✓	2020
Sigterms	SF	×	✓:GO	✓	2009
SpidermiR	Bioc	×	×	✓	2020
ToppMiR	Online	×	✓:GO	✓	2021

Table 1. Comparison of miRNA-mRNA integration tools: Many tools are R packages that can be downloaded from Bioc (Bioconductor) or SF (SourceForge). Other tools can be installed locally or are online. Few are capable of handling time series datasets. Several tools can perform funct (functional) analysis, usually using GO, Kegg, React (Reactome), DAVID or others (+) and a few tools can reduce the volume of data. Finally, this table also shows when each tool was last updated.

Many existing tools (Table 1) have particular strengths, but none satisfy the criteria necessary to bridge longitudinal multi-omic data and GRN creation. *anamiR*, *miRIntegrator*, *MAGIA2*, *Sigterms* and *SpidermiR* have substantial miRNA-mRNA integration capabilities but cannot handle longitudinal datasets (Wang *et al*., 2019; Diaz *et al*., 2017; Bisognin *et al*., 2012; Creighton *et al*., 2008; Cava *et al*., 2017). Web-based tools such as *miRNet*, *miRTarVis+* and *ToppmiR* have excellent visualisation capabilities but also cannot analyse longitudinal datasets (Fan and Xia., 2018; L'Yi *et al*., 2017; Wu *et al*., 2014). *DREM2* and *miARMa-seq* handle longitudinal datasets, but do not reduce the volume of data enough for GRN generation (Schulz *et al*., 2012; Andrés *et al*., 2016). *miRComb* can use longitudinal data to generate miRNA-mRNA interactions networks, but the networks lack detail on upstream or downstream information, making the output insufficient for GRN generation (Vila-Casadesús *et al*., 2016). Furthermore, several tools have not been actively maintained so their usability may be diminished.

There is clearly a need for a tool that can integrate, functionally analyse and generate detailed networks from longitudinal miRNA-mRNA datasets, which can then be used to identify GRNs. Here, we present the *R/ Bioconductor* package *TimiRGeN*, which uses differential expression (DE) data as input to generate small miRNA-mRNA interaction networks. Results from *TimiRGeN* can be exported to *Cytoscape* or *PathVisio* for further bioinformatic analysis (Smoot *et al*., 2011; Kutmon *et al*., 2015). The *TimiRGeN* package thereby provides a much-needed means to generate hypotheses from longitudinal multi-omic datasets. To demonstrate the capabilities of the package several datasets were analysed (see methods), including a comprehensive RNAseq time series miRNA-mRNA folic acid (FA) induced mouse kidney injury dataset (Fig.1) (Craciun *et al*., 2016; Pellegrini *et al*., 2016).

2 Methods

FA data from GSE61328 (miRNA) and GSE65267 (mRNA) were downloaded using the *fastqc-dump* function from *SRA toolkit* and fastq files were checked with *FastQC* (Leinonen *et al*., 2010; Andrews *et al*., 2010). *Cutadapt* removed adapter sequences from miRNA fastq files, and then the trimmed fastq files were processed with *mir2deep* (*mapper, quantifier* and *miRDeep2* functions) to produce mature miRNA data which could be *quant* aligned and quantified the mRNA fastq files, and *tximport* imported the output of *Salmon* into R (Patro *et al*., 2017; Soneson *et al*., 2015). Mouse transcriptome GRCm38.cdna.all was indexed for miRNA processing with *Bowtie build* and mRNA processing with *Salmon index* (Langmead *et al*., 2010; Cunningham *et al*., 2019). In R, *limma* was used for DE analysis. (Ritchie *et al*., 2015). The *makeContrasts* function performed time point based DE. The zero time point was contrasted against each subsequent time point (1, 2, 3, 7 and 14 days after folic acid injection). Results were analysed with the *TimiRGeN* R package. For the FA kidney injury dataset, the combined mode of analysis found the "Lung fibrosis" WikiPathway (WP3632) to be consistently enriched during days 3, 7 and 14 of the time course. The "Lung fibrosis" pathway was analysed for potential miRNA-mRNA interactions. Twenty interactions were kept because they were found in at least two databases and had Pearson correlations lower than -0.5. Results were exported to create a dynamic miRNA integrated Lung fibrosis signalling pathway in *PathVisio*. *CellDesigner* was then used to create a SBML formatted GRN (Funahashi *et al*., 2008). A second mouse kidney injury dataset generated by Unilateral Ureter Obstruction (UUO) was downloaded from GSE118340 (miRNA) and GSE118339 (mRNA) (Pavkovic *et al*., 2017). UUO and FA datasets were processed and analysed using the same methods. A ten time point longitudinal miRNA-mRNA breast cancer dataset was downloaded and processed as is described in the supplementary data. This dataset underwent two separate analysis with *TimiRGeN*. Once where *DESeq2* was used for pairwise DE and a second time where DESeq2 performed whole timecourse DE with the LRT method (Baran-Gale *et al*., 2016; Love *et al*., 2014). A microarray hypoxia dataset was downloaded from GSE47534 and also put through *TimiRGeN* analysis (Camps *et al*., 2014). The *lumi* and *AgiMicroRna* packages were used for processing and *limma* for pairwise DE (Du *et al*., 2008; López-Romero *et al*., 2011). Microarray platforms GPL6884 and GPL8227 were downloaded and gene IDs extracted to create a list of probes for enrichment analysis. Scripts and data for reproducibility are linked to in the supplementary data.

3 Results

3.1 Time point and microRNA specific analysis

Pairwise miRNA and mRNA DE data (Log2FC and adjusted P values) from each time point can be used as input for *TimiRGeN*. The tool works on RNAseq and microarray data, and it has two modes of analysis. The combined mode analyses miRNA and mRNA data from the same time point together, and here each gene from a time point can be filtered for significance independent of all other time points. The separate mode analyses miRNA and mRNA data independent of each other. Separate mode analysis allows for a miRNA or mRNA from a time point to be filtered for significance independent of all other time points and gene types (miRNA or mRNA). *TimiRGeN* uses WikiPathways for functional analysis, and most are curated by either entrez gene IDs or ensemble gene IDs so *TimiRGeN* provides both for the user. Neither of these annotation types can distinguish between -3p or -5p miRNAs, thus *TimiRGeN* also provides adjusted IDs, in case a miRNA-mRNA interaction network is generated with both the -3p and -5p versions of a miRNA.

3.2 Filtering data with time based functional analysis

TimiRGeN offers two functional analysis methods: time dependent pathway enrichment and temporal pathway clustering analysis. Both use the *rWikiPathways* package an API for the WikiPathways database to find signalling pathways of interest (Slenter *et al*., 2018).

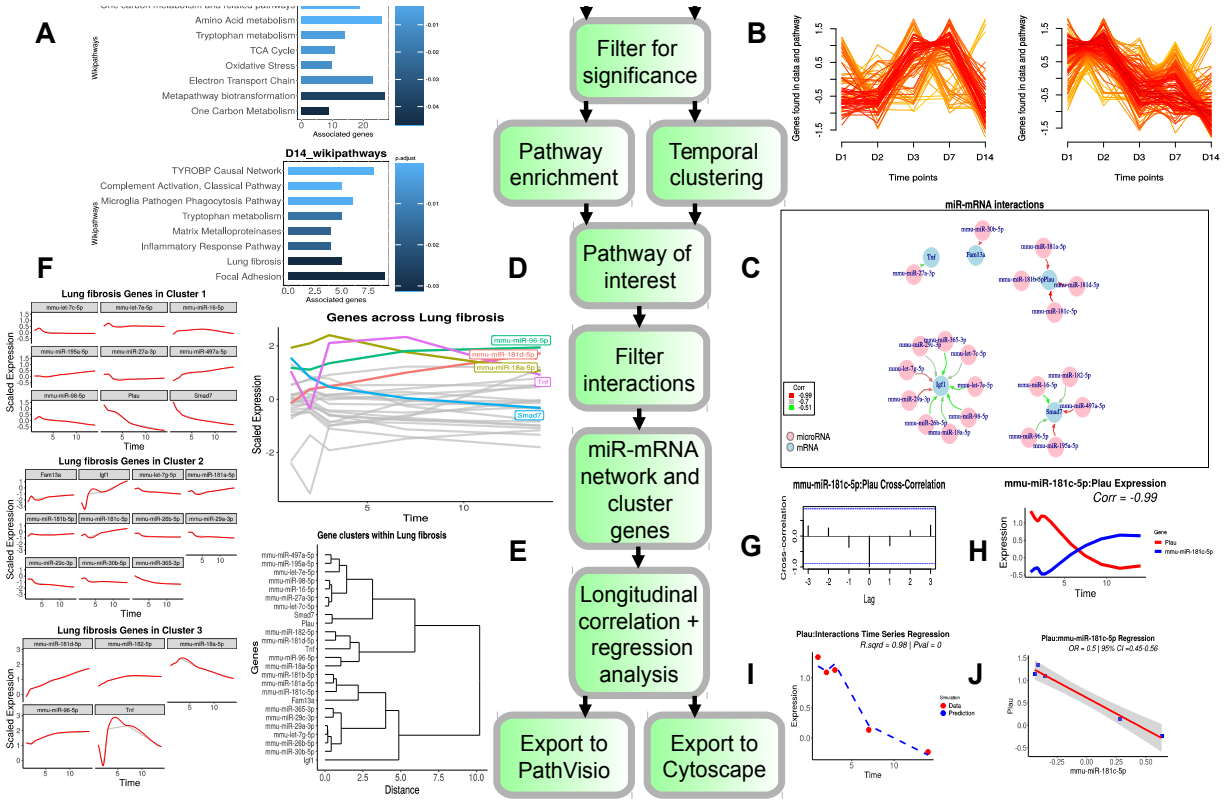


Fig. 1. Pipeline of the *TimiRGeN* R package: The FA miRNA-mRNA data are input and filtered for significantly expressed genes for each time point. From here, one of two methods can be used to find WikiPathways of interest. A) time dependent pathway enrichment to find enriched pathways at each time point. The enriched pathways are ranked in descending order of adjusted P values on bar plots. Results from day1 and day 14 are shown. Or B) temporal clustering where global trends of the pathways over time are clustered. Two clusters are shown here. Each line is a pathway and the colour represents how well a pathway fits into a cluster. Ranking from highest to lowest are: red, orange, yellow. miRNA-mRNA interactions within a selected signalling pathway can be predicted by filtration of miRNA-mRNA pairs using databases and correlation. C) Filtered miRNA-mRNA pairs can be viewed in R. Nodes are pink for miRNAs or blue for mRNAs and edges are colour coded by correlation over time. D) Behaviour of genes within the miRNA-mRNA interaction network can be viewed across the time course and genes which pass a threshold (>1.5 in this example) are highlighted. E) The genes can also be hierarchically clustered to identify trends. F) Expression changes within the clusters can be plotted. These line plots include a grey line (data points) and a red line (smooth spline). G) A selected miRNA-mRNA pair (mmu-miR-181c-5p and Plau) can be analysed using cross-correlation analysis. H) The selected mRNA (red) and miRNA (blue) can also be displayed over the time course. The data is scaled and interpolated over a spline and the correlation is displayed. I) Regression analysis can be performed on a selected miRNA or mRNA. Plau was selected, so its expression over time is predicted based on the chosen miRNAs that target it. In this example mmu-miR-181c-5p is selected to predict the behaviour of Plau. Expression values of Plau are displayed as red dots and the predicted expression of Plau is displayed as a dashed blue line. R^2 and Pvalue are shown. J) Regression can also be performed between a miRNA-mRNA pair. The OR (odds-ratio) between the two time series can be calculated, along with the 95% CI (confidence intervals). Correlation, R^2, Pvalue, OR and CI are rounded to 2 decimal places. Network data can be exported to PathVisio (S Fig.1) or Cytoscape (S Fig.2).

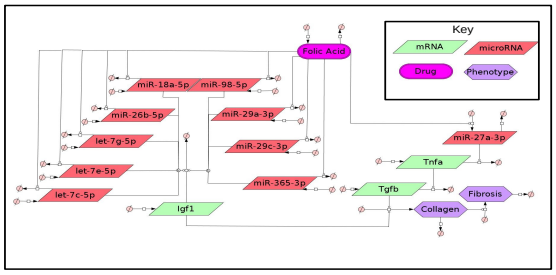


Fig. 2. miRNAs influencing anti-fibrosis factor Tnfa and pro-fibrosis factor Igf1: This GRN shows how folic acid may be down-regulating let-7c-5p, let-7e-5p, let-7g-5p, miR-18a-5p, miR-26b-5p, miR-29a-3p, miR-29c-3p, miR-365-3p and miR-98-5p, which are all predicted to target pro-fibrosis factor Igf1. Also this GRN indicates how FA may up-regulate miR-27a-3p, which is predicted to target anti-fibrosis factor Tnfa. Reduction of Tnfa will increasing levels of pro-fibrosis factor Tgfb.

Overrepresentation analysis from *clusterProfiler* is applied to time series data (Yu *et al*., 2012). Hypergeometric tests are performed to contrast the number of genes found in common between each time point (after filtering for significantly differentially expressed genes) and each species specific WikiPathway. This produces a list of enriched pathways for each time point (Fig.1A). Alternatively, if the separate mode of analysis is applied, enrichment analysis is performed for each time point per gene type. The background/ universe used to perform overrepresentation analysis can be adjusted by the user e.g. probes in a microarray or all known genes within a cell type.

3.2.2 Temporal pathway clustering method

Temporal pathway clustering (Fig.1B) utilises *Mfuzz* (Kumar *et al*., 2007). Supervised soft clusters are created based on temporal patterns which stem from the number of genes found in each time point (after filtering for significance) and each species specific WikiPathway. This will show global trends within the dataset. Pathways are assigned fitness scores for each cluster, from 0-1, and these can be filtered to find highly correlating pathways in clusters of interest. If the separate mode is used, temporal pathway clustering is performed for each gene type individually.

3.3 Filter miRNA-mRNA interactions from a signalling pathway of interest

After a signalling pathway has been selected for further analysis, the *TimiRGeN* pipeline will extract each mRNA that is found in common between the selected pathway and the input mRNA data. Each of these mRNAs are assumed to be potential targets of every miRNA in the input data. This results in a miRNA-mRNA interaction matrix which can be used to filter out miRNA-mRNA interactions that are not likely to occur by using correlations and miRNA-mRNA interaction databases TargetScans, miRDB and miRTarBase (Agarwal *et al*., 2015; Chen *et al*., 2020; Huang *et al*., 2020). Correlations are calculated between changes over time (Log2fc or average expression) between a given miRNA and a given mRNA. The default method is Pearson, but users can also select between Spearman or Kendall. Since miRNAs negatively regulate mRNAs, highly negative correlation values from miRNA-mRNA pairs could be used to identify miRNA-mRNA interactions that are likely regulate the selected pathway. Users can define a correlation threshold to filter for miRNA-mRNA interactions. The default setting for maximum correlation is -0.5. Three miRNA target databases are also usable to filter for miRNA-mRNA interactions. This includes two predictive target databases (TargetScans and miRDB) and one functional database (miRTarBase) which has had all functional support labelled as "weak" removed. Predictive databases TargetScans and miRDB were selected because, although they have differences in their prediction methods, they share usage of 3'UTR-seed site complementarity and seed site conservation to predict miRNA-mRNA interactions (Peterson *et al*., 2014). Comparisons between different miRNA-mRNA prediction methods find that 3'UTR-seed site complementarity identify the most true positive miRNA-mRNA interactions (Mazière *et al*., 2007; Zhang and Verbeek., 2010). Interactions found or not found in the three databases will be represented as 1 or 0 respectively. Users have the option to choose which combination of databases they wish to mine information from and they can choose the number of databases which an interaction needs to be mined from to be filtered. The default setting for the minimum number of databases needed to filter a miRNA-mRNA interaction is 1. Once correlation and databases have been used to filter for miRNA-mRNA interactions which may be affecting the signalling pathway of interest, they can be displayed in an internal R network (Fig.1C). Resulting genes found in the miRNA-mRNA interaction network can be viewed over the time (Fig.1D). The genes can also be sorted into hierarchical clusters shown by a dendrogram, from which clusters can be plotted to show the behaviour of the genes (Fig.1E-F). A heatmap which is compatible with the dendrogram can also be generated (S Fig.3).

3.4 Longitudinal miRNA-mRNA pair analysis

The *TimiRGeN* R package has a suite of longitudinal analysis approaches for analysing predicted miRNA-mRNA interacting pairs. This includes several correlation and regression based methods which are commonly used to analyse longitudinal datasets (Ding and Bar-Joseph., 2020). Cross-correlation analysis is a useful method to determine similarity between two time series (Fig.1G). If the time series is of sufficient length, the metric could be used to identify delays and further filter for miRNA-mRNA interacting pairs with interesting dynamics (Jung *et al*., 2011; Lakshmipathy *et al*., 2007). miRNA-mRNA pairs can also be plotted in a time series line plot. This plot can be scaled and interpolated over a spline (Fig.1H). Two types of regression analysis can also be performed. Firstly, a linear model is generated from a selected gene (mRNA or miRNA) and any number of its predicted binding partners. The combination of miRNA-mRNA interactions are left for the user to define. The longitudinal behaviour of the selected gene is predicted based on the binding partners used in the linear model. The predicted simulation and the gene data are plotted along with the R^2 value and Pvalue (Fig.1I). This type of regression prediction is useful in cases where a mRNA is targetted by multiple miRNAs or if a miRNA targets multiple mRNAs. Next, a linear model can be created from a single miRNA-mRNA pair. The odds-ratio is calculated from the regression coefficient. This measures the likelihood of one gene influencing the behaviour of another gene and has previously been used as a metric to determine miRNA-mRNA relationships (Jayaswal *et al*., 2009). 95% confidence intervals are calculated which give a range where there is a 95% certainty of the mean of the data being within the range (Fig.1J) (Szumilas., 2010). Selecting a miRNA-mRNA pair to investigate can be made easier by plotting a heatmap which orders the interacting pairs by descending correlation (S Fig.4). Statistics generated from correlation and regression analyses may be over-estimations if too few time points are found within the input data. Thus the tool will error if fewer than three time points are detected and warnings are issued if fewer than five time points are detected.

3.5 Output of the TimiRGeN package and exportation of data from R

TimiRGeN is an open-ended tool that exports to networking softwares *PathVisio* and *Cytoscape* for further *in silico* analysis. The *TimiRGeN* R package produces two data files for upload onto *PathVisio*. A file which includes a single result type, e.g. Log2FC, from each time point and gene IDs. This can be uploaded into *PathVisio* to show how the genes in a signalling pathway of interest change over the time course. Also a file which contains all filtered miRNAs can be uploaded into *PathVisio*. The second file requires the user to install the *MAPPbuilder* app in *PathVisio* (Kutmon *et al*., 2015). With this, changes over time in a miRNA integrated signalling network of interest can be visualised to show how the miRNAs may be influencing the signalling pathway. This type of display is ideal for bottom-up GRN construction (S Fig.1). Filtered miRNA-mRNA interactions can also be exported to *Cytoscape* for improved visualisation and alternative analysis via *Cytoscape* apps (Smoot *et al*., 2011). The enhanced graphics of *Cytoscape* are especially useful to visualise large numbers of miRNA-mRNA interactions (S Fig.2).

The FA kidney injury dataset had pairwise DE performed using the zero time point as the denominator. This type of pairwise analysis is recommended for time series datasets with <8 time points, however longer time series datasets may be more suitable for DE without using the pairwise approach e.g. over a cubic spline, masigpro or the LRT method with DESeq2 (Conesa *et al*., 2006; Spies *et al*., 2019). In these cases, users are recommended to filter out significantly differentially expressed genes from averaged count or expression data, and to use this as input for *TimiRGeN*. Pathway enrichment can be used to identify the most enriched pathways from the whole time course or temporal clustering can first cluster genes based on temporal behaviour. From here, genes can be sorted based on clusters, and then pathway enrichment can be used to identify enriched pathways from each temporal cluster. An alternative pipeline is shown in S Fig.5 and this is explained in section 5 of the vignette.

3.7 Datasets with multiple interventions

More complex datasets may include interventions other than time. In these cases, *TimiRGeN* should be used for each individual time series and then the results can be contrasted between different interventions. This requires the same signalling pathway to explored in each time series. As an example, the "Lung fibrosis" pathway was analysed in the FA and UUO datasets. A pipeline is shown in S Fig.6 and section 6 of the vignette provides detail for this.

3.8 Hypothesis generation with TimiRGeN

To demonstrate the tools ability to generate biologically relevant hypotheses, the FA mouse kidney injury dataset was analysed with *TimiRGeN* (Fig.1). Findings from the analysis were used to hypothesise how of FA can induce fibrosis. A GRN was constructed to formalise the hypotheses (Fig.2). Investigation of these results can be used to ratify the miRNA-mRNA interactions predicted by *TimiRGeN* and make a stronger case for experimental validation. FA injection is known to cause acute injury conditions in the kidneys, resulting in a reversible chronic kidney disease (CKD) like condition (Craciun *et al*., 2016; Pellegrini *et al*., 2016). During the 14 day time course, a number of different processes occur, such as inflammatory response, scar tissue forming, wound healing, cytokine activity (Leask and Abraham., 2004). *TimiRGeN* analysis highlights several of these processes and GRNs were generated to represent how miRNAs may be influencing fibrosis factors (Fig.2) and scar tissue forming by collagen synthesis (S Fig.7-S Fig.10). The GRN presented in Fig.2 indicates that *Igf1* acts as a miRNA sponge. Many of the presented miRNA-*Igf1* interactions have been reported, including *miR-18a, miR-26b, miR-98* and *miR-365* (Liu *et al*., 2017, 2016; Hu *et al*., 2013; Sun *et al*., 2019). *let-7c-5p* has been reported to target *Igf1*, and *TimiRGeN* predicted other let-7 family genes *let-7e-5p* and *let-7g-5p* also target *Igf1* (Liu *et al*., 2018). Finally, miR29 family members are predicted to target *Igf1*, and research indicates that *Igf1* is a *miR-29* family sponge (Gao *et al*., 2016). It is unknown why *Igf1* may be a miRNA sponge, but *Igf1* is known to induce collagen production, which contributes to kidney fibrosis and CKD (Hung *et al*., 2013). Exploration of *Igf1* as a miRNA sponge in kidney injury conditions could be beneficial for therapeutics for CKD. Overall, this case study highlights that the *TimiRGeN R* package can be used to identify biologically relevant miRNA-mRNA interactions from potentially tens-of-thousands of possible miRNA-mRNA interactions. The ability to reduce the volume of big multi-omic data is an important feature of *TimiRGeN* and one which could lead to making miRNA research easier and faster for users. Further analysis on a breast cancer dataset is also found in the supplementary data (S Fig.11-S Fig.16).

As recognised in Bar-Jones *et al* (2012), generation of more complicated transcriptomic datasets will continue, so computational biologists will need more sophisticated and up-to-date software to analyse these datasets (Bar-Joseph *et al*., 2012). Here, we have presented a novel *R/Bioconductor* package which aims to help researchers find direction when working with large longitudinal multi-omic datasets. Overall, we believe this is a useful new tool which could become a part of miRNA-mRNA data analysis pipelines.

Supplementary data

Supplementary data contains additional work. 1) Extra figures not shown in Fig.1. 2) Alternative pipelines for non pairwise DE analysis and multivariate datasets. 3) Alternative analysis of the FA kidney injury dataset. 4) A complete workflow for a breast cancer study. Including identification of a suitable dataset, processing and performing analysis with *TimiRGeN* to generate a GRN which identifies miRNAs that influence TGF-beta driven tumour fibrosis. 5) Links to *TimiRGeN R* scripts for reproducibility, vignette and a download link are also found in this file.

Funding

KP, IMC and DY are supported by the Dunhill Medical Trust [R476/0516]. DPS is supported by Novo Nordisk Fonden Challenge Programme: Harnessing the Power of Big Data to Address the Societal Challenge of Aging [NNF17OC0027812]. CP, DY and SC by the MRC and Versus Arthritis as part of the Medical Research Council Versus Arthritis Centre for Integrated Research into Musculoskeletal Ageing (CIMA) [MR/R502182/1].

Conflicts of interest

None.

References

Andrés-León, E. *et al*. (2016) miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci. Rep.*, 6, 25749.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data.

Agarwal, V. *et al*. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4, e05005.

Bar-Joseph, Z. *et al*. (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, 13(8), 552-564.

Baran-Gale, J. *et al*. (2016) An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response in MCF-7 breast cancer cells. *Rna*, 22, 1592-603.

Bartel, D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116, 281-297.

Barter, M. J. *et al*. (2015) Genome-wide MicroRNA and gene analysis of mesenchymal stem cell chondrogenesis identifies an essential role and multiple targets for miR-140-5p. *Stem cells*, 33, 3266-3280.

Bisognin, A. *et al*. (2012) MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits. *Nucleic Acids Res.*, 40, W13-W21.

Boyd, N. F. et al. (2010) Breast Tissue Composition and Susceptibility to Breast Cancer. *J Natl. Cancer Inst.*, 102, 1224-1237.

Broen, J. C. et al. (2014) The role of genetics and epigenetics in the pathogenesis of systemic sclerosis. *Nat. Rev. Rheumatol.*, 10(11), 671-681.

Camps, C. *et al*. (2014) Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia. *Mol. Cancer*, 13, 28.

Cava, C. *et al*. (2017) SpidermiR: an R/bioconductor package for integrative analysis with miRNA data. *Int. J. Mol. Sci.*, 18, 274.

in vascular smooth muscle cells from rats with kidney disease. *Plos One*, 10, e0131589.
Chen, C. *et al*. (2017) Urinary miR-21 as a potential biomarker of hypertensive kidney injury and fibrosis. *Sci. Rep.*, 7, 1-9.
Chen, Y. and Wang, X. (2020) miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.*, 48, D127-D131.
Craciun, F. L. *et al*. (2016) RNA sequencing identifies novel translational biomarkers of kidney fibrosis. *J. Am. Soc Nephrol.*, 27,1702-1713.
Cordenonsi, M. *et al*. (2011) The Hippo Transducer TAZ Confers Cancer Stem Cell-Related Traits on Breast Cancer Cells, *Cell*, 147, 759-772.
Conesa, A. *et al*. (2006) maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9), 1096-1102.
Creighton, C. J. *et al*. (2008) A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *Rna*, 14, 2290-2296.
Cunningham, F. *et al*. (2019) Ensembl 2019. *Nucleic acids Res.*, 47, D745-D751.
Cunnington, R. H. *et al*. (2014) The Ski-Zeb2-Meox2 pathway provides a novel mechanism for regulation of the cardiac myofibroblast phenotype. *J. Cell Sci.*, 127, 40-49.
Desgrosellier, J. S. and Cheresh, D. A. (2010) Integrins in cancer: biological implications and therapeutic opportunities. *Nat. Rev. Cancer*, 10, 9-22.
Diaz, D. *et al*. (2017) MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping. *Pacific symposium on biocomputing*, 390-401.
Ding, J. and Bar-Joseph, Z. (2020) Analysis of time series regulatory networks. *Curr. Opin. Syst. Biol.*
Du, P. *et al*. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13), 1547-1548.
Elston, R. and Inman, G. J. (2012) Crosstalk between p53 and TGF- Signalling. *J. Signal Transduct.*, 1-10.
Fan, Y. and Xia, J. Stechow L. V. and Delgado, S. A. (Eds.) (2018) miRNet-Functional Analysis and Visual Exploration of miRNA-Target Interactions in a Network Context. *Computational Cell Biology. Methods in Molecular Biology*, vol 1819, 215-233. Humana Press, New York, NY.
Friedlander, M. R. *et al*. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, 40, 37-52.
Funahashi, A. *et al*. (2008) CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *P. IEEE*, 96, 1254-1265.
Gao, S. *et al*. (2016) IGF1 3'UTR functions as a ceRNA in promoting angiogenesisby sponging miR-29 family in osteosarcoma. *J. Mol. Histol.*, 47, 135-143.
Genovese, F. et al . (2014) The extracellular matrix in the kidney: a source of novel non-invasive biomarkers of kidney fibrosis? *Fibrogenesis Tissue Repair*, 7.
Hanahan, D. and Weinberg, R. A. (2011) Hallmarks of Cancer: The Next Generation. *Cell*, 144, 646-674.
Heldin, C. H. et al. (2012) Regulation of EMT by TGF in cancer. *FEBS Lett.*, 586, 1959-1970.
Helwak, A. *et al*. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153, 654-665.
Hu, Y. *et al*. (2013) MicroRNA-98 induces an Alzheimer's disease-like disturbance by targeting insulin-like growth factor 1. *Neurosci. Bull.*, 29, 745-751.
Hung, C. F. *et al*. (2013) Role of IGF-1 pathway in lung fibroblast activation. *Respir. Res.*, 14, 102.
Huang, H. *et al*. (2020) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.*, 48, D148-D154.
Jayaswal, V. *et al*. (2009) Identification of microRNAs with regulatory potential using a matched microRNA-mRNA time-course data. *Nucleic Acids Res.*, 37(8), e60-e60.
Jung, D. E. *et al*. (2011) Differentially expressed microRNAs in pancreatic cancer stem cells. *Pancreas*, 40(8), 1180-1187.
Kriegel, A. J. et al. (2012) The miR-29 family: genomics, cell biology, and relevance to renal and cardiovascular injury. *Physiol. Genomics*, 44(4), 237-244.
Kumar, L. and Futschik, M. E. (2007) Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*, 2.
Kutmon, M. *et al*. (2015) PathVisio 3: an extendable pathway analysis toolbox. *Plos Comput. Biol.*, 11, e1004085.
L'Yi, S. *et al*. (2017) miRTarVis+: Web-based interactive visual analytics tool for microRNA target predictions. *Methods*, 124, 78-88.
Lakshmipathy, U. *et al*. (2007). MicroRNA expression pattern of undifferentiated and differentiated human embryonic stem cells. *Stem Cells Dev.*, 16(6), 1003-1016.
Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie2. *Nat. Methods*, 9, 357.
Laudato, S. et al. (2017) P53-induced miR-30e-5p inhibits colorectal cancer invasion and metastasis by targeting ITGA6 and ITGB1. *Int. J. Cancer*, 141, 1879-1890.
FASEB J., 18, 816-827.
Leidinger, P. *et al*. (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, 14.
Leinonen, R. *et al*. (2010) The sequence read archive. *Nucleic Acids Res.*, 39, D19-D21.
Levental, K. R. et al. (2009) Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling. *Cell*, 139, 891-906.
Liu, C. *et al*. (2017) miR-18a induces myotubes atrophy by down-regulating IgfI. *Int. J. Biochem Cell Biol.*, 90, 145-154.
Liu, F. and Di Wang, X. (2019) miR-150-5p represses TP53 tumor suppressor gene to promote proliferation of colon adenocarcinoma. *Sci Rep.*, 9.
Liu, G.-X. *et al*. (2018) Hsa-let-7c controls the committed differentiation of IGF1-treated mesenchymal stem cells derived from dental pulps by targeting IGF-1R via the MAPK pathways. *Exp. Mol. Med.*, 50.
Liu, H. *et al*. (2016) MicroRNA-26b is upregulated in a double transgenic mouse model of Alzheimer's disease and promotes the expression of amyloid-B by targeting insulin-like growth factor 1. *Mol. Med. Rep.*, 13, 2809-2814.
Liu, T. *et al*. (2019) Cancer-associated fibroblasts: An emerging target of anti-cancer immunotherapy. *J .Hematol. Oncol.*, 12, 86.
López-Romero, P. (2011). Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics*, 12(1), 1-8.
Love, M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12), 1-21.
Lu, P. et al. (2012) The extracellular matrix: a dynamic niche in cancer progression. *J. Cell Biol.*,196, 395-406.
Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, 17, 10-12.
Mazière, P and Enright, A. (2007) Prediction of microRNA targets. *Drug Discov.*, 12, 452-458.
Miyaki, S. *et al*. (2010) MicroRNA-140 plays dual roles in both cartilage development and homeostasis. *Genes Dev.*, 24, 1173-1185.
Meng, X. M. et al. (2016) TGF-B: The master regulator of fibrosis. *Nat. Rev. Nephrol.*, 12, 325-338.
Moore, M. J. *et al*. (2015) miRNA-target chimeras reveal miRNA 3-end pairing as a major determinant of Argonaute target specificity. *Nat. Commun.*, 6, 1-17.
Ooi, C. Y. *et al*. (2018) Network modeling of microRNA-mRNA interactions in neuroblastoma tumorigenesis identifies miR-204 as a direct inhibitor of MYCN. *Cancer Res.*, 78, 3122-3134.
Patro, R. *et al*. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14, 417-419.
Pavkovic M. *et al*. (2017) Multi omics analysis of fibrotic kidneys in two mouse models. *Sci. Data*, 6, 92.
Peterson, S. M. *et al*. (2014) Common features of microRNA target prediction tools. *Front. Genet.*, 5, 23.
Pellegrini, K. L. *et al*. (2016) Application of small RNA sequencing to identify microRNAs in acute kidney injury and fibrosis. *Toxicol. Appl. Pharmacol.*, 312, 42-52.
Principe, D. R. *et al*. (2014) TGF-B: duality of function between tumor prevention and carcinogenesis. *J. Natl. Cancer I.*, 106.
Proctor, C. J. *et al*. (2017) Computer simulation models as a tool to investigate the role of microRNAs in osteoarthritis. *Plos One*, 12.
Provenzano, P. P. *et al*. (2009) Matrix density-induced mechanoregulation of breast cell phenotype, signaling and gene expression through a FAK-ERK linkage. *Oncogene*, 28, 4326-4343.
Ritchie, M. E. *et al*. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43, e47-e47.
Roche, J. (2018) The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers*, 10.
Roush, S. and Slack, F. J. (2008) The let-7 family of microRNAs. *Trends Cell Biol.*, 18(10), 505-516.
Schwarzenbach, H. *et al*. (2014) Clinical relevance of circulating cell-free microRNAs in cancer. *Nat. Rev. Clin. Oncol.*, 11.
Schulz, M. H. *et al*. (2012) DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst. Biol.*, 6.
Selbach, M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455, 58-63.
Slenter, D. N. *et al*. (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, 46, D661-D667.
Smoot, M. E. *et al*. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27, 431-432.
Soneson, C. *et al*. (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000research*, 4.
Spies, D. *et al*. (2019) Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Brief. Bioinform.*, 20, 288-298.

induced acute kidney injury and early fibrosis. *Toxicol. Lett.*, 224, 326-332.
Su, B. et al. (2014) Let-7d suppresses growth, metastasis, and tumor macrophage infiltration in renal cell carcinoma by targeting COL3A1 and CCL7. *Mol. Cancer*, 13, 206.
Sun, W. *et al*. (2019) miR-365 inhibits duck myoblast proliferation by targeting IGF-I via PI3K/Akt pathway. *Biosci. Rep.*, 39.
Szumilas, M. (2010) Explaining odds ratios. *J. Am. Acad. Child Adolesc. Psychiatry*, 19(3), 227.
Tang, C. M. et al. (2017) CircRNA 000203 enhances the expression of fibrosis-associated genes by derepressing targets of miR-26b-5p, Col1a2 and CTGF in cardiac fibroblasts. *Sci. Rep.*, 7, 1-9.
Qin, S. *et al*. (2015) Gene regulatory networks by transcription factors and microRNAs in breast cancer. *Bioinformatics*, 31, 76-83.
Vila-Casadesús, M. *et al*. (2016) MiRComb: an R package to analyse miRNA-mRNA interactions. Examples across five digestive cancers. *Plos One*, 11.
Wang, J. P. and Hielscher, A. (2017) Fibronectin: How its aberrant expression in tumors may improve therapeutic targeting. *J. Cancer*, 8, 674-682.
Wang, T.-T. *et al*. (2019) anamiR: integrated analysis of MicroRNA and gene expression profiling. *BMC Bioinformatics*, 20.
Wang, R. *et al*. (2019) Long noncoding RNA DNM3OS promotes prostate stromal cells transformation via the miR-29a/29b/COL3A1 and miR-361/TGF1 axes. *Aging*, 11, 94429460.
Wen, X. *et al*. (2012) One dose of cyclosporine A is protective at initiation of folic acid-induced acute kidney injury in mice. *Nephrol. Dial. Transplant.*, 27, 3100-3109.
Wu, C. *et al*. (2014) ToppMiR: ranking microRNAs and their mRNA targets based on biological functions and context. *Nucleic Acids Res.*, 42, W10-W113.
Yao, L. *et al*. (2019) Paracrine signalling during ZEB1-mediated epithelialmesenchymal transition augments local myofibroblast differentiation in lung fibrosis. *Cell Death Differ.*, 26, 943-957.
Yu, G. *et al*. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, 16, 284-287.
Zaha, D. C. (2014) Significance of immunohistochemistry in breast cancer. *World J. Clin. Oncol.*, 5, 3820392.
Zhang, Y. and Verbeek, F. J. (2010) Comparison and Integration of Target Prediction Algorithms for microRNA Studies. *J. Integr. Bioinform.*, 7, 169-181.

8.2 Appendix - B. Poster presentations

All presentations have been uploaded onto this github repository https://github.com/Krutik6/
Thesis_Data_Scripts/tree/main/PhD_presentations. They can be found by adding the website extension which is in the Link column of Table 8.1.

Event	Link
Systems Biology: Making Sense of Complexity	/London_2018.pdf
ICSB2018	/Lyon_2018.pdf
EpiGenOA2018	/Dublin_2018.pdf
ICSB2019	/Okinawa_2019.pdf
BSU meeting	/Newcastle(BSU)_2020.pdf
ISGSB2020	/SouthAfrica_2020.pdf
A&G theme meeting	/Newcastle(A%26G).pdf
Journal Club	/Newcastle(JournalClub)_2021.pdf
Bioc2021	/US_2021.pdf

Table 8.1: **List of presentation performed as conferences or workshops.** Through the last four years I have delivered three poster presentations at conferences/ workshops and four oral presentations, and this included being invited as a speaker at Bioc2021. I have also delivered an oral presentation at our group journal club.

8.3 Appendix - C. Scripts and data

Most code and data needed to reproduce work presented in this Thesis is readily available from https://github.com/Krutik6/Thesis_Data_Scripts/tree/main/Data. Some data has not been made accessible because it has been generated by collaborators and they have selected not to make the data public. Also, the validation data has not been made available because we seek to publish this work.

Ch2

Code and data needed to reproduce work presented in Ch2 is stored in https://github.com/Krutik6/TimiRGeN/issues/1. A README file explains the contents of this repository. The *TimiRGeN* R package is found in https://github.com/Krutik6/TimiRGeN. It can be installed in R using the following commands:

$>$ library(devtools)

$>$ install_github("Krutik6/TimiRGeN")

or using *BioCManager*

$>$ BiocManager::install("TimiRGeN")

The vignette for the package is linked in the *Bioconductor* repository - http://www.bioconductor .org/packages/devel/bioc/vignettes/TimiRGeN/inst/doc/TimiRGeN_tutorial.html

Ch3

https://github.com/Krutik6/Thesis_Data_Scripts/tree/main/Data_Ch3 contains four folders: DE, pathvisio, Preprocessing and TimiRGeN. Raw and normalised data have not been uploaded in this repository, as my collaborators did not publish their raw/normalised data [113]. Preprocessing and DE folders contain scripts for miRNA and mRNA data. Pairwise DE input for analysis with TimiRGeN has been made available and a script for this can be found in the corresponding folder. Some TimiRGeN output is available in the pathvisio file. This is used to create dynamic signalling pathways.

Ch4

https://github.com/Krutik6/Thesis_Data_Scripts/tree/main/Data_Ch4 contains two folder are found: Modelling and Rplotting. The former contains the COPASI kinetic model, along with all information about species, reactions, functions and ODEs and experiment files 1 and 2. Experiment file 1 is used for parameter estimating to the calibration data. Experiment file 2 is not used to parameter estimation, but rather as a barometer to see how the behaviours should change during *miR-199b-5p* inhibition. Rplotting contains a script and COPASI output files to create plots.

Ch5

https://github.com/Krutik6/Thesis_Data_Scripts/tree/main/Data_Ch5 contains data and code used in this section is available. Age, gender and Qvalue files contain data and scripts for DE analysis, and this also contains the QC work e.g. PCA plots. Raw and Normalized counts for all 336 sequencing experiments are stored here. Feature engineering work and ML work is found here in the respective folders. Preprocessing folder contains material used to miRNAseq and RNAseq processing.